Theory and Utility—Key Themes in Evidence-Based Assessment:
Comment on the Special Section

Richard M. McFall
Indiana University—Bloomington

This article focuses on two key themes in the four featured reviews on evidence-based assessment. The first theme is the essential role of theory in psychological assessment. An overview of this complex, multilayered role is presented. The second theme is the need for a common metric with which to gauge the utility of specific psychological tests and measures for specific purposes. A metric from information theory is recommended. The implications of these themes for the four reviews and for the future of psychological assessment in general are discussed.

Keywords: evaluating assessment, measurement theory, measurement utility, incremental validity, quantifying information value

The special section on developing guidelines for the evidence-based assessment of adult disorders published in this issue of Psychological Assessment treats us to four scholarly articles by prominent authorities who critically review the state of the art of psychological assessment in four distinct problem areas. Two focus on Axis I disorders identified in the Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM-IV; American Psychiatric Association, 1994) that are prevalent in adults, namely, forms of anxiety (Antony & Rowa, 2005) and depression (Joiner, Walker, Pettit, Perez, & Cukrowicz, 2005); one focuses on a major Axis II category, personality disorders (Widiger & Samuel, 2005); and another focuses on a coding issue facing the editors of the fifth edition of the DSM, namely, couple distress (Snyder, Heyman, & Haynes, 2005). The first three deal primarily with the assessment of individuals, mostly relying on off-the-shelf, standardized, self-report measures, whereas the fourth deals with the assessment of dyadic relationships, mostly through coding of behavioral observations. The articles by Antony and Rowa (2005), Joiner et al. (2005), and Snyder et al. (2005) tend to mix studies addressing nomothetic questions (i.e., using assessments to test theories about the nature and etiology of a problem) and studies addressing idiographic questions (i.e., using assessments in the diagnosis and treatment of individual cases). The article by Widiger and Samuel (2005) provides a more focused review of basic research addressing primarily nomothetic questions. Each raises a number of thought-provoking conceptual and methodological issues.

Not being an expert in any of these problem areas but having a long-standing interest in psychological assessment more generally, I welcome these reviews as definitive and enlightening tutorials. They clearly should be relevant to investigators and practitioners specializing in these problems but should be of interest to a much broader audience as well, because they highlight several common themes that are of fundamental importance to all psychological assessment. The reviewers’ emphasis on their focal topics kept them from covering the broader themes in depth. Therefore, this commentary expands on two key themes in the reviews and explores their implications for psychological assessment in general. Although the reviews tend to emphasize established psychological tests, my comments are intended to be generic, applying broadly to psychological assessment. The aim is to promote empirically supported assessment.

Theme 1: The Role of Theory
in Psychological Assessment

Too often, the essential role of theory in psychological assessment is overlooked. Fortunately, this was not the case in the current special section. The authors of the four reviews within this section raised fundamental questions about their core theoretical constructs: anxiety, depression, personality disorders, and couple distress. At a general level, they asked how well these constructs carve nature at its joints and contribute to our understanding of relationships among key variables. At a more specific level, they asked whether these clinical problems were best conceptualized within the DSM system, whether a categorical or a dimensional framework fits the data best, how to handle the sticky issues of comorbidity and multidimensionality, and how to deal with other thorny conceptual and methodological issues inherent in each of these clinical problems. This kind of frank, skeptical treatment of cherished theoretical constructs is all too rare but is essential to scientific progress.

A thorough evaluation of psychological assessment in any problem area should start with a critical examination of the quality of the overarching theoretical and measurement models. Just as water cannot rise naturally above its source, and just as a measure’s reliability sets an upper limit on its potential validity, the utility of a psychological test or measure is limited by the quality of the theoretical and measurement models it serves. To the extent that such models do a poor job of representing nature, any tests and

I am indebted to Teresa A. Treat, Yale University, for her helpful comments on drafts of this article.

Correspondence concerning this article should be addressed to Richard M. McFall, Department of Psychology, Indiana University, Bloomington, Indiana 47405-1301. E-mail: mcfall@indiana.edu
measures linked to the models necessarily will be of limited value. In this spirit, the authors of the four reviews presented in the special section remind us repeatedly that it is meaningless to talk about the reliability, validity, and utility of specific tests or measures apart from their theoretical context and intended purposes. Measures never are ends in themselves but have value only to the extent that they have nomothetic or idiographic utility—that is, to the extent that they shed new light on theoretical questions or improve the accuracy and efficiency of practical decision making in specific situations.

How does this emphasis on theory fit with the fact that some of the most successful instruments in psychology’s toolbox are based on an atheoretical, actuarial approach? Does not this success discredit the claimed importance of theory in assessment? I would argue that the answer is no. The actuarial approach is atheoretical only in its method of selecting and aggregating test items to create predictor variables. For example, the Minnesota Multiphasic Personality Inventory (MMPI), a poster child for this genre, is a measure of psychopathology, focusing on specific categories of pathology, all of which are theoretical constructs. When test items are selected actuarially, without regard to the face validity or theoretical relevance of their content, this actuarial strategy rests on a big theoretical assumption—namely, that the conditional probabilities observed in the past between responses to test items, on one hand, and criterion measures, on the other, will hold true in the future. This theoretical notion, which is relatively new in the history of science (Gigerenzer et al., 1989), has been fruitful but is not foolproof. If one relied on it to pick stocks on the New York Stock Exchange, for example, one could go broke. But the most serious limitation of the actuarial approach is that it contributes little or nothing to our understanding of why test responses and criterion variables are related. Insurance actuaries may be forgiven for not asking “why?” when predicting risk and setting rates, but science is concerned with explanation as well as with prediction. Without investigating the basis of relationships, we are vulnerable to predictive errors stemming from such problems as the “third variable problem” and the “error of the assumed essence.” That is, we are prone to assume that variables are related causally when, in fact, their association is merely an artifact of shared influence from one or more unexamined variables.

Whereas the actuarial approach is neither theory free nor without limitations, its predictive success to date makes it a useful benchmark against which to compare the results of more elaborate theoretical approaches to assessment and prediction. Just as placebos and minimal treatments provide a standard of comparison in treatment research, actuarial tests provide a baseline standard of parsimony against which to measure the incremental validity of other, more complex theories and methods in assessment. We will return to this issue of incremental validity under the second theme.

This first theme—the importance of theory—reminds us that psychological tests and measures are but one part of a complex, multilayered, logical, bootstrapping structure by which scientists build, test, refine, and apply explanatory and predictive models of natural phenomena. To appreciate the unique contributions of psychological tests and measures, as well as their limitations, one needs a clear schematic map of the multiple layers of this conceptual structure, the functions served at each layer, and how the different layers are interrelated. The following heuristic framework (adapted from McFall & Townsend, 1998) provides such an overview.1

At the most abstract level of this structure are the postulates (philosophical assumptions, myths, values, beliefs, metaphors), which are the untestable “givens” that form the exoskeletal framework for any theoretical model (Polya, 1957; J. M. Smith, 1984). For instance, anxiety, depression, personality disorders, and couple distress are valued negatively as “problems.” In the present reviews, three problems are viewed metaphorically as “mental illnesses,” or “deviant behaviors” within individuals, whereas the couple distress problem is viewed as a “system problem” involving “dysfunctional” or “maladaptive” interactions between individuals. Such postulates influence where one looks for answers. Of course, one might just as easily start from different postulates, perhaps focusing on environmental influences, such as expressed emotion, on psychopathology (e.g., Butzlaff & Hooley, 1998) or focusing on the role of individual typologies in couple distress and violence (e.g., Holzworth-Munroe & Stuart, 1994). Indeed, any problem can be viewed from any number of perspectives—social, cognitive, biochemical, genetic, economic, political, and so forth. There is no one right way to approach a problem. The key question, which cannot be answered with confidence until reaching the final level in this structure, is, “What does it buy me to view this problem from this perspective?” (See Johnson, 1946; Miller, 1996.)

At the next lower level of the structure are the formal theoretical constructions—that is, hypothetical constructs and their nomological networks, formal relationships among different constructs, and hypothesized functions and processes that these constructs supposedly serve or regulate (Cronbach & Meehl, 1955; MacCorquodale & Meehl, 1948). It is a common logical error to accept uncritically one’s pet hypothetical constructs—to reify them, treating them as though they were “real things”—rather than treating them for what they are: made-up, conjectural abstractions designed to capture regularities, or recurrent patterns, in experience. Although no two events ever are exactly alike, the function of hypothetical constructs is to emphasize the similarities in the stream of unique events while minimizing the differences. Anxiety, depression, personality disorder, and couple distress are prominent examples of such formal hypothetical constructs.

When a theory’s formal constructions are specified with clarity and precision, they provide a seed bed for germinating testable hypotheses and predictions. Useful scientific models not only are able to account for the events that gave rise to them in the first place but also can sprout new, risky predictions about novel events. The scientific worth of formal theoretical propositions ultimately is judged on the basis of how well they increase our ability to predict events that we could not have predicted without them (Feynman, 1985; Popper, 1962). From this perspective, the distance separating theoretical and applied assessment shrinks; these are interdependent and inseparable aspects of an organic process.

While hypothetical constructs and the regularities they represent are the building blocks of theoretical models, individual constructs cannot be evaluated in isolation any more than one can understand

---

1 The framework owes much to J. B. Rotter’s lectures on assessment at The Ohio State University in the 1960s (see also Rotter, 1954).
world geography by cutting out the dots representing cities on the map and examining them in an isolated, disconnected way. Individual constructs are subcomponents within a larger theoretical system, with specified relationships to other constructs, so they must be evaluated in terms of how well they cohere with the other elements of this larger system. A construct gains explanatory power (i.e., has construct validity) to the extent that it functions as expected within this system. Not only must the nodes comprising the construct’s nomological net relate strongly to one another as predicted, but they also must relate (or not relate) as expected to the nodes of other constructs in the larger system. Things that are supposed to go together theoretically should go together empirically; things that are not supposed to go together should not go together.

For example, different measures of the construct of anxiety (i.e., the measures at the nodes in the nomological net for anxiety) are expected to be positively and strongly associated with one another and positively, but less strongly, related to measures of related neighboring constructs (e.g., stress, arousal, fear, avoidance). These same measures are expected to be relatively independent of theoretically distinct constructs (e.g., depression). In short, they should demonstrate both convergent and discriminant validity (Campbell & Fiske, 1959). Unfortunately, there is no conventional metric, no threshold similar to the .05 level of a significance test, for deciding when a pattern of results is sufficiently close to the expected pattern to be judged to have convergent and discriminant validity. Too often, when there is a big gap between expected and observed patterns, investigators simply focus on confirming evidence, ignoring the disconfirming evidence. For instance, “comorbidity” sometimes is used as a fancy excuse for embarrassingly poor discriminant validity. Indeed, given the evident comorbidity between anxiety and depression, these constructs increasingly are viewed as related disorders. In that case, however, their underlying theoretical and measurement models need to be modified accordingly. Does it even make sense to continue to regard them as separate disorders rather than as facets of a common disorder?

The levels of the structure covered thus far have been very abstract. But for scientists to test the power of their theoretical models, they ultimately must find ways to instantiate them. Only then will they be able to quantify the degree of correspondence between their theoretical predictions and actual observable events. HEREIN lies a problem, however: Hypothetical constructs are unseen forces, or organizing principles, posited as underlying explanations for perceived regularities in events. They cannot be observed directly; they only can be examined indirectly by studying samples of discrete events that are considered imperfect and incomplete reflections of them.

The empirical process of testing a theory’s formal propositions, then, starts at the next lower level of the structure—the referent level. Referents are the category labels for the observable exemplars located at the nodes in the nomological nets that define the abstract hypothetical constructs and their relationships. Referents for the construct of anxiety, for example, might include subjective distress, autonomic arousal, or escape and avoidance behaviors. (See Lang, 1968, and Kozak & Miller, 1982, for detailed discussions of the problems with such referents in a tripartite model of fear and anxiety.) In formal psychological tests, referents sometimes are referred to as test constructs. Examples would be Vocabulary or Digit Span on the Wechsler Adult Intelligence Scale—

Third Edition (Wechsler, 1997). Referents are abstractions themselves, and still must be tied down specifically to events that can be sampled and recorded. For example, if we selected autonomic arousal as the referent for anxiety, we still would need to decide how to sample and record such arousal. That occurs at the next level of the model.

Where do referents come from, and how do assessors choose among the possibilities? Referents can come from anywhere. The source is irrelevant; ultimately, all that matters is how well a referent performs. But all other things being equal, when the conceptual links between referents and their constructs are direct and obvious, this facilitates the logical interpretation of results. Unfortunately, there is no sure way of knowing in advance how well a referent will reflect its underlying construct. All choices among potential referents are conjectures, leaps of faith, influenced as much by lay language, metaphor, and hunch as by any scientific observation, theory, or algorithm. If we asked a layperson to tell us how we would know if someone were anxious (or depressed, or suffering from a personality disorder or from an unhappy relationship), the layperson’s list of indices (referents) probably would mirror those offered by psychologists for the same construct. This is not surprising, given that most psychological constructs are adopted from lay language (Mischel, 1968). Psychologists’ primary contributions have been to refine and systematize these lay concepts and to improve the methods by which referents are defined, sampled, quantified, and evaluated.

Referents come in two main varieties: signs and samples. A sign is any event that is believed to have a statistically predictive relationship to the criterion of interest. A sign’s content need not be related logically to the criterion; the only requirement is that it be predictive. Projective tests and objective, actuarial tests typically take a sign approach. Signs can be treated either categorically (one response is sufficient to classify someone) or additively (the more signs, the stronger the characteristic). Item response theory is a prominent system for determining which items, in which combinations, should be included in a given test or measure (also see G. T. Smith, Fischer, & Fister, 2003). Over the years, clinical psychologists have shown a preference for the sign approach (in contrast to cognitive, developmental, neurologic, and sensory psychologists). Whatever the sign approach’s appeal, it is not based on any clear evidence of predictive superiority over the sample approach.

In the sample approach, content is relevant. Referents are selected specifically because they are believed to replicate critical aspects of the criterion. Sample responses are presumed to represent the responses that would occur in the criterion situation. The responses may be sampled either directly through performance measures, or indirectly through retrospective self-reports (e.g., daily diaries, estimates of alcohol consumption per unit of time) or reports by collaterals (e.g., parents, teachers, spouses, friends).

---

2 The term sample is used in two ways here, which could be confusing. A sample always is a collection of representative events, but in the sample approach, the term emphasizes the importance of maximizing the similarity between one’s sample assessment tasks and responses, on one hand, and one’s criterion tasks and responses, on the other hand. This is in contrast to the less restrictive samples requirements for tasks and responses in the sign approach.
Some common examples of performance-based sample tests are psychomotor tests (e.g., typing, driving, or dexterity tests), cognitive tests (e.g., recognition memory, recall, or classification tests), problem-solving tests (e.g., mazes, vocabulary, logic, or math tests), social skills tests (e.g., social role plays and simulations), and creativity tests (e.g., writing, music, and drawing samples). All such tests are designed to provide representative samples of the very activities they are designed to predict. IQ tests, which arguably are the most successful of all psychological tests, can be viewed as sample tests, systematically sampling the multiple theoretical aspects of intellectual performance and predicting performance across a range of related criteria.

In general, sample measures are founded on two related theoretical assumptions: (a) that the best predictor of future behavior in any given situation is past behavior in that same situation and (b) that the more faithfully the assessment task captures the essential features of the criterion task, the more accurately it will predict the criterion. Thus, performance-based measures (when they are feasible) usually are considered preferable to self- or collateral-report measures because they provide the most direct samples with the best verisimilitude. Identifying and simulating essential aspects of the real-world criterion requires a thorough analysis of the criterion task, its subcomponents, and the task-relevant behaviors that determine success. The choice of referents in the sample approach typically is guided not only by theory but also by analytic and analogical reasoning.

In the present set of reviews, the article on couple distress (Snyder et al., 2005) provided the most extensive coverage of the sample approach to assessment, with attention to some of its strengths and weaknesses. In general, researchers interested in interpersonal behavior problems—from children’s social development to social influences on aging and health—have relied heavily on measures that sample social interactions through role plays, simulations, and structured tasks designed to assess typical or optimal performance. The goal in these social measures is to identify distinctive features of adaptive versus dysfunctional interactions that may explain the origins of the interpersonal problems while also having implications for the design of remedial interventions.

Ironically, assessors who take a sample approach, and who pay close attention to issues of representativeness, tend to be criticized when their assessment and criterion tasks are less than isomorphic. Investigators of couple distress, for instance, have been criticized for using laboratory simulations of couple interactions that necessarily are artificial; the results of such measures are dismissed as unrepresentative of couples’ natural interactions at home. Meanwhile, the critics may be using self-report measures with items that bear little resemblance to the content of their criterion. This is an interesting asymmetry. Attempt to sample criterion content in a highly representative way, and any discrepancy is grounds for criticism; make no such effort, and major discrepancies are not a serious problem.

On closer examination, the sign versus sample distinction should not be treated in a black-and-white manner. All assessments may be arrayed along a continuum reflecting the isomorphism between the measure and the criterion. At the sign end, for example, MMPI items were selected actuarially, but many were written with an eye toward capturing theoretically relevant aspects of specific categories of psychopathology. At the sample end, laboratory simulations designed to capture couples’ naturalistic inter-actions may be artificial, but if they provide predictive information about a couple’s relationship, they still may be useful. Across the continuum, what ultimately matters is a measure’s utility for predicting things that could not have been predicted otherwise.

By what implicit or explicit criteria—other than a box score of predictive accuracy—is it possible to evaluate how well one’s referents actually reflect one’s superordinate constructs? The her-editary writings on construct validity (e.g., Campbell & Fiske, 1959; Cronbach & Meehl, 1955) have described the ideal ways that referents for a hypothetical construct should relate to the construct and to one another; however, empirical results from the trenches seldom live up to these ideals. So, how close must the results be for referents to be judged as representative? Unfortunately, there are no absolute rules or general conventions. When referents do not behave as expected, other factors may be involved, such as the way the referents were operationalized, sampled, quantified, analyzed, or interpreted (all to be discussed below) rather than to deficiencies in the referents themselves. Before dismissing referents prematurely, other possible reasons for any shortcomings need to be ruled out first. There simply is no sure way to determine whether one’s choice of referents is good without proceeding through the remaining layers of the structure.

Thus, we have arrived, at last, at the level of instrumental methods—that is, tests, instruments, tasks, techniques, and procedures. This is the level at which actual behavior is sampled, the level most people have in mind when they think about psychological assessment. This is the most concrete level of the structure thus far, the level at which a theory’s hypothetical constructs and referents are wrestled to earth and made tangible through the power of operational definitions. Indeed, some psychologists act as though psychological assessment starts and ends at this level, treating instrumental methods as ends in themselves. These assess-sors uncritically adopt instrumental methods “off the shelf,” merely assuming that a given test’s name accurately summarizes what it measures. Furthermore, they assume that if a test is standardized, has published research on its psychometric properties, and is widely used, then it must be valid for their specific purpose and situation. If the test is commercially published, this seems to inspire the utmost confidence. Such uncritical acceptance of instrumental methods is unwarranted, of course. The choice of methods should be dictated by the local circumstances, including one’s theoretical constructs and referents, the particular assessment purpose, the specific evidence, and the respective costs and benefits of the available options.

Choices among methods seldom are cut-and-dry. If autonomic arousal were our referent for the theoretical construct of anxiety, for example, it might be sampled in a variety of ways—from recording electrodermal conductivity, heart rate, or brain activity with psychophysiological instruments to recording subjective levels of arousal with structured interviews or paper-and-pencil, self-report measures to sampling muscle twitches or fidgeting via behavioral observations. All choices among potential instrumental methods are conjectures as to what will work best. Even if a method has been shown empirically to be useful for one context or purpose, this does not mean it will be useful for this new context or purpose. A healthy skepticism is needed: “Is this method appropriate here? Is there a better choice?” But it always is more
defensible and less arbitrary to choose the method with the strongest empirical track record in similar circumstances.

No single method provides a definitive sample of either the hypothetical construct or the referent. Assessors often seem to lose sight of this fact, as reflected in their erroneous tendency to equate answers on an IQ test with intelligence, increased skin conductance with arousal, self-reports of lethargy with depression, and so forth. Instrumental methods are important, but their role in the overall scheme must be kept in perspective: They are simply the tools for gathering samples of events that are presumed to reflect indirectly the theoretical constructs and referents of interest.

Several concerns come into play at the level of instrumental methods. The first is the need to control contamination of the sample by unintended influences, including all the various threats to internal validity (Campbell & Stanley, 1963). Using standardized methods is a big first step toward reducing such contamination, as is anticipating possible threats and designing controls for them.

A second concern is the repeatability or reliability of measurement. Customary ways of estimating reliability (e.g., test–retest, split-half, item total, interjudge, Cronbach’s alpha) should be familiar. However, two common arguments relating to reliability and validity need clarifying. First, it often is argued that it always is best to obtain the largest sample possible, as the results will be more reliable and, hence, more valid. This argument usually is tied to the logic of the familiar Spearman–Brown prophecy formula. Second, in a similar vein, it often is argued that it is best to use multiple methods to assess a referent, rather than relying on just one, on the grounds that the pooled estimates of the target variable will be more reliable and, hence, more valid.

To help clarify the issues behind these two related arguments, I suggest considering them from a Bayesian perspective. According to this point of view, sample-size requirements are determined by (a) the discrepancy between one’s a priori estimates and actual observations; (b) the inherent variability of the phenomenon being measured; and (c) the error of one’s measurement method. One should continue sampling until the resulting estimates of the sample parameters have stabilized; that is, until additional cases no longer significantly decrease the discrepancy between the iterative a priori and a posteriori estimates. To continue sampling beyond this point of diminishing returns is wasteful.

In an extreme case, if you want to know someone’s name, age, sex, address, and telephone number, it is reasonable to assume from past experience that there would be little gain in reliability or stability from asking for this information multiple times (i.e., by increasing the sample size or by using multiple methods). Whereas single-item samples may be adequate for some variables, other variables may require extremely large samples and multiple methods before they yield stable parameter estimates. Even after parameter estimates have stabilized, other factors—that is, the inherent variability of the phenomenon or inadequacies of the method—may result in large errors of prediction, with wide confidence intervals. Complicating matters further, if a variable is inherently unstable, simple parameter estimates may do a poor job of capturing the phenomenon, no matter how large one’s sample. Other ways of modeling the phenomenon quantitatively may be needed. The point is that while larger samples generally yield more reliable estimates, decisions about sample size should be guided by rational considerations and empirical results rather than by rigid rules.

What about the use of multiple methods? Garb (2003) has addressed this issue. Quite simply, more is not always better. Adding the results of less reliable and valid measures to the results of a reliable and valid measure is not likely to increase predictive accuracy. Indeed, adding the results from garbage measures, simply for the sake of using multiple methods, will pollute one’s results. The predictive utility of specific combinations of measures cannot be assumed but must be demonstrated empirically for each assessment task.

A third concern highlights a common problem. It is a concern with the internal logic of the instrumental methods. That is, are the sampling methods, on one hand, and the theoretical referents that the methods were designed to represent, on the other, logically consistent or compatible? An example of this logic problem would be the use of self-report questionnaires to assess individual differences in men’s sensitivity to women’s heterosxual cues. If one’s theory suggests that sexually coercive men are insensitive to women’s cues, then it seems logically inconsistent to expect insensitive men to report accurately on their own sensitivity. In another example, if one’s theory proposes that individuals with a personality disorder are devious, then it seems illogical to give credence to their self-reported truthfulness. In a third example, it seems illogical to include highly similar questions on two different instruments, each supposedly tapping the referents for a distinct construct, and then to treat positive correlations between the two measures as evidence that one construct explains the other. In a final example, it seems illogical to use a computer touch screen to assess individual differences in response time, as a measure of cognitive processing, in persons with and without Huntington’s disease. Given that Huntington’s disease impairs motor control, such a measure cannot provide an uncontaminated assessment of cognitive processing.

A final concern is a common failure to appreciate the distinctions between instrumental methods that are designed as “scales” and those designed as “inventories.” Scales are more common than inventories. All items comprising a scale are assumed to tap a common construct, although constructs may be decomposed further into two or more component factors or facets. Typically, scale items simply are added up rather than given differential weight; essentially, this means that the items are treated as interchangeable. Scale items, particularly within factors, are expected to be highly correlated. As G. T. Smith et al. (2003) have pointed out, however, items should be chosen to represent a construct’s entire domain, and every item should add to the whole picture. If items are too highly correlated, they will be redundant, adding little new information to the representation of the overall construct, even though they might improve the reliability of the estimates of the individual facets comprising the construct domain.

In contrast to scales, inventories typically take a sampling approach, collecting responses to a heterogeneous set of relatively independent, minimally overlapping items, each contributing uniquely to the probabilistic prediction of the target outcome. Generally, inventories are designed to predict outcomes, not to assess a single construct. The underlying assumptions are that outcomes have multiple determinants, and that the contribution of any specific determinant to the outcome in a particular occasion is complex and probabilistic. Thus, inventory items are designed to sample broadly from a logically coherent set of task-specific elements that make up the complex matrix of determinants, each
contributing only probabilistically to the outcome of interest. The predictive value of any given item depends probabilistically on the match between the task demands actually faced by a person (or system) on a given occasion; the specific capacity of the person (or system) to deal effectively with those demands, as assessed by relevant items; and the likely consequences of failing to manage the demands (i.e., sometimes the failure to manage a task has no important consequences). Inventory items are not viewed as interchangeable; by design, they map nonoverlapping areas. They also should not be highly correlated, and should not fit neatly into factor analytic solutions. Given that there always is a practical limit on the number of possible items, there inevitably is a trade-off between breadth and depth of item coverage. While increased sample size should increase reliability, each item devoted to this purpose is an item that could have been devoted to increasing the absolute number or density of areas covered. To the extent that item scores on an inventory are treated additively, this simply indicates that the covered areas are treated as having an equal likelihood of affecting the outcome. If a total inventory score is used to predict an outcome, this is based on a risk assumption; that is, the more areas a person or system is unable to manage effectively, the higher the risk of encountering a situation that causes a problem. An alternative to using total scores is to use profiles to make conditional predictions of specific task performance.

Although inventory methods are not as common as scales in psychological assessment, when they have been developed and evaluated for problems ranging from delinquency, to depression, to eating disorders, they have yielded very promising results (e.g., Fisher-Beckfield & McFall, 1982; Freedman, Rosenthal, Donahoe, Schlundt, & McFall, 1978; Gaffney & McFall, 1981; McFall, Eason, Edmondson, & Treat, 1999; Ward & McFall, 1986).

Because inventories are less familiar, an analogy may help to illuminate their logic. Suppose you were organizing your backpack for a wilderness camping trip. You would begin by making a list of the things you needed to pack. This packing list would be designed to maximize both its fit to the anticipated demands of the trip and its efficiency (pack light!). Each item should serve some important, independent function (avoid redundancy!). Items would not be interchangeable; omitted items would leave a void. Given such item independence, the probability of one item’s presence on the inventory list would reveal little about the probability of any other items being on the list. To the extent that we, as outside observers, have general knowledge about what events are likely to be encountered on wilderness camping trips, we should be able to generate an inventory of items that should be included in your backpack. Furthermore, by comparing what you actually packed to this inventory list, we should be able to make reasonably accurate probabilistic predictions about how well prepared you are for your camping trip. Indeed, we could predict which specific situations (if encountered) you will or will not be prepared to handle effectively.

In general, to make predictions using the inventory approach, one needs to know four things: (a) what critical events are likely to be encountered and their probability distributions; (b) what resources are required to deal effectively with these situations; (c) what resources actually are available for dealing with each of the probable events; and (d) the likely consequences of not being prepared to deal with specific events should they occur, that is, the impact of inventory gaps. Information in the first, second, and fourth categories comes first, from a careful task analysis; information in the third category comes from an inventory assessment, which is linked to the prior task analysis. Of course, any predictions necessarily must be stated as conditional probabilities. That is, outcomes are conditional on what events occur; the inventory’s profile of available resources becomes predictive only in those situations where resources actually are required. In our camping analogy, failing to pack a snake-bite kit could be fatal but only in the low-probability event of a poisonous snake bite. Failure to pack waterproof matches could cause significant problems but only in the moderately likely event that one’s regular matches get wet. However, failure to provide adequately for potable water requirements almost certainly will cause serious problems very quickly.

A few examples of the inventory approaches to assessment are the Binet, designed to assess French children’s readiness for the demands of French public schools; the Office of Strategic Services effort, in World War II, to select candidates for the spy service through simulated performance in situations that spies might face after being dropped behind enemy lines; modern military selection and training of fighter pilots; or the National Football League’s (NFL) selection of athletes on the basis of performance samples across a range of athletic tasks at the NFL Combine. The inventory approach is used routinely in personnel psychology. When hiring for a position, the first step is to conduct a job analysis, listing the expected responsibilities and competencies of the position. The next step is to assess systematically the job candidates’ capacities to fulfill these expectations. Rather than relying on a summary score on a single dimension, each candidate’s profile of qualifications typically is matched to the profile of the job requirements.

All of the concerns related to the development of instrumental methods illustrate the complexities involved in adequately representing the referents for our constructs. Unfortunately, even if we managed to address all these concerns, we still would not be able to determine with confidence whether our constructs, referents, and methods were successful until we had worked our way through the remainder of the structure.

Once we have gathered samples with our methods, we must decide what to do with them. Thus, the next level of the structure involves the measurement model, where the sampled referents are converted into units and assigned numerical values on some scale. The goal, of course, is to model the sampled events quantitatively in ways that preserve their essential qualities, such as order and magnitude. The distinction between methods and measures is important. The sample gathered by a method may be used to generate any number of potential measures. The responses to a paper-and-pencil test, for example, could yield measures of time to completion, pressure on the pencil, number of “true” responses, and so forth. At the same time, there may be multiple methods for generating a given type of measure. For instance, a measure of vocabulary level could be obtained with a variety of different methods. When assessing the referent of length, rulers and yardsticks are instrumental methods, whereas centimeters and inches are measures.

If the referent for anxiety were autonomic arousal, for example, and if arousal were sampled with psychophysiological recordings of the electrodermal conductivity associated with palmar sweat gland activity, then our measurement model might represent this activity as micro-ohm units on an ordinal or interval scale over a specified time, grouped into sampling periods of a given bin size. If the referent were subjective arousal, then the measurement
model might represent this on an ordinal or interval scale of the sum of keyed self-report responses to a fixed number of specific items on a paper-and-pencil questionnaire. And if the referent were avoidance behavior, the measurement model might represent this on nominal or ordinal scales of standardized coding categories for observers’ judgments about target behaviors in a structured task over specified periods. However, all such decisions are creative acts of faith; referents, methods, and measures do not come with owners’ manuals specifying how they should be quantified and what measurement scales are best.

The reasoned specification of measurement scales (nominal, ordinal, interval, ratio) in measurement models is critical to the validity of researchers’ inferences about their measurements. The fundamental scale requirements of scientific measurement models, along with their associated statistical, inferential, and theoretical implications, have been developed elsewhere in detail (e.g., Krantz, Luce, Suppes, & Tversky, 1971; Roberts, 1979; Townsend, 1990; Townsend & Ashby, 1983, 1984). Unfortunately, clinical researchers frequently disregard these measurement dictums. Perhaps the most egregious example is the treatment of ordinal measurements as if they were interval measurements. For example, Likert-scale responses and total scores on self-report dictums. Perhaps the most egregious example is the treatment of ordinal measurements as if they were interval measurements. For example, Likert-scale responses and total scores on self-report do not readily accessible. However, with contemporary analytical strategies are dictated by the basic assumptions?

Each analytic method carries with it assumptions about the essence or “average” of a data set (mean, median, mode?). Should we focus on level, frequency, period, amplitude, slope, intercept, or what? Should we aggregate over time, or keep time as a factor? Are things stable or changing over time? What is the best bin size for collapsing data into units? Should we treat data categorically or dimensionally? With each decision, potentially valuable information in the data may be lost, or noise may be mistaken for signal.

The next-to-last level involves data analysis—that is, where the data and summary statistics are processed further by statistical methods, mathematical models, oracular tests, and so forth. As Meehl (1971, 1978) pointed out, there is no such thing as an “automatic inference machine.” One’s choice of data analytic techniques should be determined by one’s theory, measurement model, and specific questions. All choices at this level must be congruent with all the choices at higher levels of this structure. Each analytic method carries with it assumptions about the essential character of the phenomena represented by the data, assumptions that may, or may not, be compatible with the overarching theoretical model or with the phenomena being measured. For example, analyzing transactional data with, say, time-series analyses, rather than with the interaction term in an analysis of variance (ANOVA), may be appropriate for some questions but not others.

Similarly, whether it is appropriate to attempt to control for so-called nuisance variables depends, first of all, on the theoretical model being tested (see Meehl, 1971). Even if such control seems appropriate theoretically, it may not be feasible logically and technically. For example, analysis of covariance (ANCOVA) frequently is used in psychopathology research to “control” for initial group differences on potential covariates; however, Miller and Chapman (2001) have explained why this almost always is inappropriate, yielding meaningless results. Essentially, if two groups differ initially on a covariate, no statistical technique can change that reality, artificially making the groups equal, without yielding make-believe nonsense. Quite simply, statistical analyses are informative only to the extent that they address theoretical questions appropriately.

If the data were being used to address an abstract, nomothetic question (e.g., comparing two research samples to test a theory about the attributes associated with personality disorder), the analytical strategy might be very different than if the data were being used to explore a more applied nomothetic question (e.g., comparing two treatments in a clinical trial to decide which is more cost-effective). The strategy for analyzing data in an individual clinical case is likely to be quite different from that in either of the nomothetic examples. Analytic strategies are dictated by the questions.

Only on reaching the final level of the structure—the interpretation and inference level—is it possible, at long last, to evaluate what, if anything, has been learned by going through the seven preceding steps in this multilayered structure. What, if anything, was learned from (7) the data analyses of (6) the summary statistics (5) generated by the measurement model (4) of the responses gathered by the instrumental methods (3) designed to sample the referents (2) for the formal theoretical constructs (1) supported by the basic assumptions?

Answering this question requires some point of reference—a criterion, or “gold standard”—against which to judge the relative information value and implications of the assessment results. However, the criterion must be measured quantitatively before it can serve this function. This requires that the same multilayer structure, outlined above, be applied to the assessment of the criterion. That is, what are the postulates, constructs, referents, instrumental methods, measurement models, summary statistics, data analytic methods, and interpretations and inferences that define the criterion? Think of this parallel criterion-assessment process, then, as assessment squared. Only on reaching the final level of the structure—the interpretation and inference level—is it possible, at long last, to evaluate what, if anything, has been learned by going through the seven preceding steps in this multilayered structure. What, if anything, was learned from (7) the data analyses of (6) the summary statistics (5) generated by the measurement model (4) of the responses gathered by the instrumental methods (3) designed to sample the referents (2) for the formal theoretical constructs (1) supported by the basic assumptions?
Decisions about the validity of assessments inevitably depend on the choice of criteria. The same test results might be judged as valid when viewed in relation to one criterion but invalid when viewed against a different criterion. The same criterion measure also might produce different results depending on the characteristics of the sample (age, sex, ethnicity, etc.) or assessment arrangements. This means that validity cannot be considered an intrinsic property of one’s instrumental methods or measures, per se, but must be viewed as conditional on all of the particular assessment circumstances.

The interpretive and inferential process at this final stage should feed back correctly through all layers of the structure, on both sides of the equation (except, perhaps, at the postulate levels). When results line up with expectations at all higher levels, particularly on the test and measurement side, this tends to increase the stature of the overarching theoretical model and all that flows from it, but it does not constitute definitive proof of anything. The structure has so many free parameters—auxiliary assumptions, ancillary hypotheses, unconstrained decisions, and nonstandard methods—that few competing explanations can be ruled out, and no firm causal links can be established across all the multiple levels. When results are inconsistent with expectations, this may decrease the stature of the overall theory, but negative evidence almost never is sufficiently definitive to warrant tossing out the entire theory. There usually is wiggle room; again, one always can argue that an unexpected negative outcome was due to a mistaken auxiliary assumption, ancillary hypothesis, strategic decision, or choice of methods. Besides, it always is possible that any problems were due to inadequacies on the criterion side of the equation, rather than on the assessment side.

The strength of this multilayered, twin-towered structure is no greater than its weakest link, and the corrosive influence of errors tends to be multiplicative. Thus, interpretations of even the most carefully planned and executed assessments in controlled nomothetic research will be open to challenge. Idiographic interpretations, in turn, depend heavily on support from their nomothetic foundations. The only long-term strategy for addressing the inherent limitations of psychological assessments is replication. Unfortunately, replication was not a strong feature of the research covered in the featured reviews. Nor did the research discussed in these four areas seem to have a clear cumulative quality or coherence. The reviewers, of course, are not responsible for this state of affairs; they were merely the messengers. The sheer difficulty of falsifying psychological theories (Popper, 1962) and of drawing strong inferences from such research evidence (Platt, 1964) is what led Meehl (1978) to call psychology a “soft” science.

Theme 2: Evaluating Utility in Psychological Assessment

As described above, psychological assessment plays a specific, limited, but crucial role in the multilayered, twin-towered process that culminates in scientific interpretations and inferences—however soft these conclusions may be. Assessment plays this role across the full range of specific purposes and contexts—from formal hypothesis testing in a basic research context to practical decision making regarding individual cases in an applied clinical context. Whether an assessment has utility depends not only on the question being asked but also on important logical and technical aspects of the inference process itself.

In the simplest case, for instance, basic researchers focusing on nomothetic questions regularly use assessment data to decide whether to reject the null hypothesis. If the difference between groups, relative to the variability within groups, exceeds what would be expected by chance, then by applying a simple decision rule, typically based on the 95% confidence level, investigators feel comfortable rejecting the hypothesis that the groups are not different. Note the odd double negative in this statement. The logical and technical foundations of this simple, dichotomous inferential process have been challenged, however (e.g., Cohen, 1994; Gigerenzer et al., 1989; Loftus, 1996; Meehl, 1978). Logically, a decision to reject the null hypothesis is not equivalent to deciding to accept any particular alternative hypothesis; it simply means that the groups are judged to differ beyond what would be expected by chance alone, without indicating how or why. Technically, the likelihood of rejecting the null hypothesis when it is false is affected by sample size, or statistical power, which makes the likelihood of deciding correctly that groups differ unduly dependent on the investigator’s diligence.

Interpretive and inferential processes become far more complicated when assessors interested in idiographic questions use psychological tests and measures to guide practical decisions that have real-life implications for individuals or couples. For example, Is this individual anxious, depressed, or suffering from a personality disorder? or, Is this couple distressed? If so, is the problem serious enough to warrant some kind of intervention? Are reasonable interventions available for this problem? If so, which intervention is most appropriate in this case? And if an intervention has been undertaken, what effect is it having, if any? All of these questions involve logical and technical issues that go far beyond those involved in the simpler decision about whether to reject the null hypothesis.

Whether assessment data are useful in answering such idiographic clinical questions depends, among other things, on base rates, cutting scores, and inverse probabilities (Meehl & Rosen, 1955). First, variability in base rates, or prevalence, affects most traditional indices of accuracy, such as percent correct, positive predictive power, and negative predictive power. Although the indices of sensitivity and specificity are not affected directly by base rates, they are influenced indirectly because they are affected by the choice of cutting scores, and the optimal choice of cutting scores is affected by prevalence. Clinical problems seldom have population base rates over 5%, which makes it very difficult for any psychological test to contribute to increased accuracy for decision making in individual cases. If a problem has a 5% base rate, for instance, one can classify cases correctly 95% of the time without ever using a test, simply by going with the base rates and calling all cases “negative.” For a test to have utility for improving overall accuracy, it would have to beat this base rate accuracy of 95%! Psychopathology researchers may finesse this base rate problem in nomothetic research by artificially increasing the base rate—for example, by comparing clinical and control samples of equal size, or by studying high-risk samples. In idiographic assessment, however, where the goal is to make predictions about an individual case, the base rate problem cannot be sidestepped so easily.

Second, when using test scores to classify cases, one must select a cutting score—that is, the score used as the dividing line between positive and negative cases. In clinical situations, cutting scores...
often are dictated by resources, which set limits on selection ratios. For instance, it may be that only a fixed number of cases can be accepted into a treatment program at any one time. However, differences in cutting scores will yield differences in the evidence regarding the utility of a given assessment package. The choice of cutting score affects all of the conventional indices of overall accuracy, including percent correct, sensitivity, specificity, positive predictive power, and negative predictive power. A conservative choice will yield fewer false positives but also will yield fewer hits or true positives. A liberal choice, in contrast, will yield more hits but also more false positives. Thus, the choice of cutting scores invariably involves trade-offs in the types and magnitude of errors experienced. This means that the results yielded by a test are not solely an inherent characteristic of the test, per se.

A third complication is the “inverse probability problem.” This is when assessors confuse \( p(S \mid D) \), which is the probability of a particular score on a diagnostic test given membership in a diagnostic group, with \( p(D \mid S) \), which is the probability of being a member of the diagnostic group given a particular score on the diagnostic test. Estimates typically are available for the first probability from nomothetic research, but seldom for the second, which is what is needed when dealing with individual cases. Only where the base rate is 50% are these two probabilities equal, but such symmetry is rare with clinical disorders. The greater the asymmetry in base rates, the greater the inequality. Bayes’s theorem (Bayes, 1763, as cited in Gigerenzer & Murray, 1987) can minimize this problem by controlling for base rates while using normative probability distributions to estimate inverse probabilities (Meehl & Rosen, 1955). For some reason, few assessors have adopted the Bayesian solution, despite repeated warnings over many years about this underlying problem.

These three complications—base rates, cutting scores, and inverse probabilities—illustrate some of the reasons why it is so difficult to come up with a global evaluation of a given method’s utility for assessing an idiographic clinical problem. Any such evaluation not only is dependent on the specific context and purpose but also is influenced significantly by these complicating factors. Unfortunately, none of the featured reviews covered these factors in detail when examining the validity of the assessment methods commonly used in their problem areas. Whereas it may be unrealistic to look for a single, global index of a method’s utility for all applications, it may be worthwhile to look for a general approach to evaluating the relative utility of methods and measures at the local level, when used for specific purposes, with specific clinical problems.

I suggest starting by restating the point made at the beginning: the common purpose of all psychological tests and measures is to serve the two interdependent and inseparable themes of theory and utility. That is, the function of assessment is to generate information that sheds new light on specific theoretical questions and increases the accuracy or efficiency of practical decision making.

**Incremental validity** is the term used to represent the degree to which a given test or measure genuinely serves these purposes in a particular context. This term has been around at least since Sechrest’s (1963) article by the same name (see also Meehl, 1959). However, the term too often is invoked in a vague, idealized way rather than as a clearly defined standard with teeth. To be useful as a tool for advancing empirically supported psychological assessment, the concept of incremental validity needs to become a yardstick with practical value for evaluating and comparing the relative value of specific tests and measures, for specific purposes, in specific contexts. Note that for an assessment to have utility, it should have incremental validity, but assessments that have incremental validity may not have utility. To evaluate utility, however, one must understand incremental validity.

Fortunately, in December 2003 *Psychological Assessment* published a special section devoted specifically to examining incremental validity, with eight articles covering topics spanning an introductory overview (Hunsley, 2003); the history, varied meanings, design issues, and practical implications of incremental validity (Hunsley & Meyer, 2003); considerations involved in evaluating different dimensions of incremental validity in new assessment methods (Haynes & Lench, 2003); five principles for enhancing the incremental validity of a test during its construction (G. T. Smith et al., 2003); detailed ways of determining the costs, benefits, cost-effectiveness, and cost–benefit ratio of assessments (Yates & Taub, 2003); and reviews of the incremental validity of assessments in three specific applied areas: children and adolescents (Johnston & Murray, 2003), adult psychopathology (Garb, 2003), and treatment (Nelson-Gray, 2003). These excellent articles covered these topics in rich detail, offering clear and penetrating analyses, and providing a wealth of illuminating advice and wisdom. The special section on incremental validity should be required reading for all psychological assessors, whether basic or applied, working in labs or clinics.

Although there is currently no universally accepted, standard metric for quantifying a measure’s incremental validity, hierarchical multiple regression analysis is probably the most common analytic strategy for quantifying the incremental contributions of specific methods, items, or measures to existing assessments. Hunsley and Meyer (2003) have pointed out some of the inherent problems with this strategy, such as the fact that the results depend on the order in which variables are entered into the equation, with early entries earning credit for variance shared with later entries, including any nonreplicable variance associated with sample-dependent error. Even if one safely navigates the shoals of this analytic strategy and emerges with a quantified estimate of a measure’s incremental contribution, however, it is unclear how best to interpret the meaningfulness of any such increase.

McFall and Treat (1999) have suggested a metric with which it may be possible both to quantify rigorously the incremental validity, or informational value, of specific assessment methods in specific applications and to compare the informational values of different methods (or combinations of methods) across applications. Such a metric would avoid some of the limitations of the hierarchical multiple regression approach, permitting direct “apples-to-apples” comparisons of the relative information value of specific assessments. This promising metric comes from information theory.

Information theory (e.g., Pierce, 1980; Schmitt, 1969; Shannon & Weaver, 1949) explicitly defines information in terms of a reduction of uncertainty, which, in turn, is defined quantitatively as the magnitude of the decreased error resulting from a signal input to a system (or less commonly, but more positively, as the magnitude of the increased accuracy yielded by that input). This definition of information always implies a comparison of error rates under two or more conditions—at minimum, decisions made with and without a signal’s input. Viewed in this information
theory framework, then, a psychological test’s incremental validity is equivalent to the information value of its signal input to a decision system. Information value can be quantified as the increment in accuracy resulting from test information, relative to what could be achieved without the test, either by chance alone or by the use of alternative measures.

Fortunately, signal detection theory (SDT) provides a formal, well-established, quantitative system for conceptualizing and measuring the information theory concepts of information, reduction of uncertainty, decreased error, and increased accuracy (see Green & Swets, 1974; Macmillan & Creelman, 2005; McFall & Treat, 1999; Swets, 1996). Assessment systems typically are used in diagnostic decision making to enhance our ability to discriminate between two mutually exclusive states, namely, between the presence or absence of a signal. Just as a radiologist may use X-rays to decide whether a tumor is present, a psychologist may use assessments to decide whether a couple is distressed or whether a patient is suffering from anxiety, depression, or personality disorder. SDT avoids some of the confounding complications, discussed earlier, by partitioning the variability in diagnostic data into two independent components: perceptual and decisional. The perceptual component represents the system’s sensitivity, ability to discriminate, or accuracy; the decisional component represents the system’s bias, or choice of decisional boundaries or cutting points, from liberal to conservative.

By analyzing a given measure’s “receiver operating characteristics” for multiple criteria, or cutting points, it is possible to plot a curve that connects the dots for the paired hits and false-alarm rates at all cutting points. This, in turn, allows one to compute the area under the curve (AUC) for a given measure. AUC is unaffected by base rates; it is not limited to any specific cut point; and avoids the confusion of the inverse probability problem. AUC is a logical, clearly interpretable index of a measure’s incremental validity relative to chance, providing an index of the probability that pairs of observations drawn at random from two underlying distributions will be ordered, or classified, correctly. Thus, AUC provides a common scale on which the accuracy of different measures can be compared directly. A more detailed discussion of SDT is beyond the scope of this commentary, but interested readers are encouraged to pursue the subject by starting with the overviews cited above and continuing with the clinical examples cited below.

By adopting the concepts of information theory and the related analytical methods, such as SDT and its logical extensions (e.g., general recognition theory: Ashby & Townsend, 1986), investigators interested in practical problems of psychological assessment would be armed with standard, rigorous metrics for quantifying, comparing, and evaluating the information value, or incremental validity, of specific measures, used in specific ways, for specific purposes. The power and applicability of this strategy already has been demonstrated in evaluations of specific assessment methods for a variety of clinical problems, from anxiety, panic, and depression (Mossman & Somoza, 1989; Somoza & Mossman, 1992; Somoza, Steer, Beck, & Clark, 1994), to violent behavior (Mossman, 1994; Rice & Harris, 1995), to schizophrenia (Olin, John, & Mednick, 1995). Ironically, even though psychologists have played a key role in developing this approach to diagnosis and decision making, the approach thus far has gained wider acceptance in fields outside of psychology (see Swets, 1996). Perhaps this will change if psychological assessors pay more attention to the themes raised in the current reviews and elaborated in this commentary.

Implications for the Future of Psychological Assessment

The development of more coherent theoretical models and more rigorous approaches to evaluating the utility of psychological tests and measures by quantifying their incremental validity will go a long way toward promoting the use of empirically supported psychological assessments, in general, and toward advancing research on specific psychological problems, such as those reviewed in the present special section, in particular. On the basis of the information presented in the featured reviews, all four research areas appear to suffer from a number of common shortcomings. To the extent that these areas are typical of clinical research more generally, they help spotlight some very specific ways that research on clinical assessment could be improved in the future.

At the top of my list of improvements would be an increased attention to the theoretical foundations of the psychological assessment, with a decreased tendency to treat tests and measures as ends in themselves, characterized by stable, intrinsic, context-free levels of reliability, validity, and utility. In addition, it seems critical that future work become less scattered and more cumulative. This would require researchers to reach a consensus about which questions define the cutting edge of their area. The current lack of effort devoted to replication is very disquieting for all the reasons discussed here. Equally troublesome is the blurring of important distinctions, such as the different requirements for nomothetic and idiographic assessment. The most striking deficit in the research, however, is the general neglect of two crucial bottom-line issues: the problems inherent in criterion assessment and the problems involved in quantifying incremental validity. This summary list of problems in the four areas is not a criticism of the four featured reviews, each of which has made the most of the available literature. On the contrary, it is a general reflection on the inherent difficulty of critically reviewing the research literature on evidence-based assessment in a particular area when (a) investigators fail to articulate clearly the theoretical questions being addressed and (b) research results are not presented in a form that permits meaningful evaluations of the relative utility of different assessment approaches to specific clinical problems.

References


Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161–171.


approaches to assessing couple distress. Psychological Assessment, 17, 288–311.

New Editor Appointed, 2007–2012

The Publications and Communications (P&C) Board of the American Psychological Association announces the appointment of a new editor for a 6-year term beginning in 2007. As of January 1, 2006, manuscripts should be directed as follows:

- Emotion (www.apa.org/journals/emo.html), Elizabeth A. Phelps, PhD, Department of Psychology, New York University, 6 Washington Place, Room 863, New York, NY 10003.

Electronic manuscript submission. As of January 1, 2006, manuscripts should be submitted electronically via the journal’s Manuscript Submission Portal (see the Web site listed above). Authors who are unable to do so should correspond with the editor’s office about alternatives.

Manuscript submission patterns make the precise date of completion of the 2006 volumes uncertain. The current editors, Richard J. Davidson, PhD, and Klaus R. Scherer, PhD, will receive and consider manuscripts through December 31, 2005. Should 2006 volumes be completed before that date, manuscripts will be redirected to the new editor for consideration in 2007 volume.