Psychological Assessment
2005, Vol. 17, No. 4, 396–408

1040-3590/05/$12.00 DOI: 10.1037/1040-3590.17.4.396

On Construct Validity: Issues of Method and Measurement

Gregory T. Smith
University of Kentucky

Fifty years ago, L. J. Cronbach and P. E. Meehl (1955) advocated for the concept of construct validity, noting that psychologists study hypothetical, inferred entities and that validating measures of such entities involves basic theory testing. Three important developments in clinical assessment following that seminal article are noteworthy. First, clinical research has benefited from greater theoretical integration and subsequent differentiation among related constructs. Second, implementation of ongoing, critical evaluation of all aspects of the construct validity process, including theory development, hypothesis specification, research design, and empirical evaluation, has improved clinical assessment. Third, improvement in evaluating fit between hypotheses and observations has been sought. Improved means of evaluating multitrait, multimethod designs, and ways to increase their clinical representativeness, are one encouraging development. Ongoing efforts to improve the construct validity process reflect the legacy of L. J. Cronbach and P. E. Meehl.

Keywords: construct validity, theory testing, clinical research advances

This year marks the 50th anniversary of the publication of Cronbach and Meehl's (1955) classic article, "Construct Validity in Psychological Tests." The occasion is a good one to consider the impact of the concept of construct validity on clinical assessment. In what follows, I briefly review the concept, note how it has evolved over the last 50 years, suggest that it has facilitated certain advances in clinical research, and further explicate its role in clinical assessment. My intention is to facilitate improvement in the construct validation process, as researchers continue to develop new theories and measures to accompany them.

It is widely appreciated that the notion of construct validity, when first advanced, represented a significant departure from the prevailing views of the time. Those views were perhaps best summarized by Anastasi's (1950) statement that "It is only as a measure of a specifically defined criterion that a test can be objectively validated at all .... To claim that a test measures anything over and above its criterion is pure speculation" (p. 67). The dramatic contrast of the construct validity perspective, which allowed for theoretical statements concerning unobserved psychological phenomena and means for validating them, is quite apparent. In the decades since then, construct validity has not only been widely accepted, it has come to be seen as an umbrella term, describing a process for theory validation that subsumes specific test validation operations (Landy, 1986; Messick, 1980). Psychological science, and its clinical arm, has matured with the recognition that use of psychological measures represents an aspect of theory testing. The result for clinical assessment has been a clearer, more coherent framework for understanding psychological dysfunction, as well as more sophisticated, useful tools for evaluating the success of construct validation efforts.

This article is organized as follows. First, I provide an overview of the concept of construct validity. In subsequent sections, I offer an update on advances in philosophy of science, argue that researchers' understanding of construct validity has evolved productively, provide an updated model for the construct validation process, and consider concrete examples of the impact of construct validity on clinical assessment. I then offer a critical discussion of advances in evaluating empirical evidence for construct validity.

Correspondence concerning this article should be addressed to Gregory T. Smith, Department of Psychology, University of Kentucky, Lexington, KY 40506-0044. E-mail: gsmith@uky.edu

Overview of Construct Validity

Applying the classic perspective of Cronbach and Meehl (1955), psychological constructs are, essentially, unobservable. One cannot directly observe neuroticism, extraversion, dependency, or any other inferred trait. Physical science has an International Bureau of Weights and Measures with, for example, a bar reflecting the true length of a meter. Measuring length, for physicists, has an agreed-on, concrete anchor. Psychology has no such thing. We infer the existence of traits such as neuroticism because doing so has obvious utility for describing persons, their differences from each other, and the nature of dysfunction (Goldberg, 1995). We consider it important to study them because of their potential for understanding and explaining a great deal of human behavior.

Therefore, the first challenge for scientific psychology concerns how to measure hypothetical constructs such as these in a convincing, valid way. Cronbach and Meehl (1955) argued that to do so, one must demonstrate that one's measure of a given construct relates to measures of other constructs in theoretically predictable ways. For hypothetical constructs, there is no good way to determine whether a measure reflects the construct validly, except to examine whether scores on the measure conform to a theory, of which the target construct is a part. To oversimplify, if I develop a measure of hypothetical construct A, I can only validate my measure if I have some theoretical argument that, for instance, A relates positively to B, but is unrelated to C. If I have such a theory, and if I have measures of constructs B and C, I can test whether my measure of A performs as predicted by my theory. The indeterminacy of any such set of tests is apparent. If my hypothesis that A

relates to B but not C is not supported, I face many possibilities. Perhaps my theory is correct, but my new measure of A is inadequate. Perhaps my theory is correct, but the measure of either B or C is inadequate. Perhaps each measure is adequate, but my theory is fully or partially incorrect. Perhaps my theory and measures are adequate, but the design of my study contains flaws or limitations. Perhaps both my theory and my measure are inadequate. On the other hand, if my hypothesis is supported, I am still not certain I have validly measured A. Perhaps my new measure of A inadvertently overlaps with B (known not to correlate with C), and my supportive results are really due to the measures of A and B partly reflecting the same construct. There are, of course, many such possibilities.

Psychology cannot have an "international bureau of psychological constructs"; we measure inferred constructs, and the validity of any measure is part and parcel of the validity of the theory that led to the measure. Cronbach and Meehl (1955) recognized this problem and so talked about the need for bootstrapping. When one cannot begin with either proven theory or certain measurement, one must conduct a series of studies to examine different theoretical and measurement possibilities. During that process, repeated evidence consistent with the same hypothesis increases confidence in that hypothesis, even though a hypothesis is never fully proven. Thus, multiple tests of construct validity, using different criteria assessed in different ways, is a normal part of the process. Similarly, Campbell and Fiske (1959) emphasized the importance of measuring hypothetical constructs using different methods: They recognized that shared method variance accounted for substantial overlap among psychological measures, and they provided a means to assess the validity of measures above and beyond shared method variance (the multitrait, multimethod matrix [MTMM]). A defining feature of all of this work was appropriate skepticism. A certain basic skepticism is inherent in recognizing the provisional nature of our constructed representation of hypothetical psychological entities (Campbell, 1995; Fiske, 1995; Shrout, 1995).

It is important to remember that Cronbach and Meehl's (1955) emphasis was not on recording a few successfully predicted correlations. Because construct validation involved basic theory testing, they emphasized principles for making inferences about the meaning of test scores or experimental outcomes. Since their early contribution, methodologists have periodically sought to remind investigators of this crucial perspective. Messick (1980) and Guion and Cranny (1982) emphasized the design characteristics necessary for validity studies, including data-gathering procedures, the choice of variables to study, and appropriate inferences to be drawn. Lawshe (1985) reinforced this point by urging researchers to think of types of validity *analyses*, rather than types of validity. By *validity analyses*, he appeared to refer to the whole process of drawing sound inferences from empirical investigations. The validation process, in his view, should be understood as a system involving sound research design, appropriate data analysis, and suitable inferences from one's findings. Landy (1986) encouraged adoption of Lawshe's (1985) perspective, noting its similarity to original descriptions of the concept. He argued that doing so would help prevent us from viewing construct validation as collecting a series of stamps: a content validity correlation, a criterion-related validity correlation, and so on. Cronbach (1985; cited in Landy, 1986) also emphasized inferences, rather than tests.

In the clinical literature, a number of recent articles on validity reflect this emphasis on the theory-based, inferential nature of construct validity (Clark & Watson, 1995; Foster & Cone, 1995; Smith, Fischer, & Fister, 2003; Smith & McCarthy, 1995). In a book honoring Donald Fiske (Shrout & Fiske, 1995), several of the original protagonists reiterated concerns about the indeterminacy of the validation process and hence the need for careful attention to each step from theory to observation (Campbell, 1995; Cronbach, 1995; Fiske, 1995; Meehl, 1995). Interestingly, Cronbach and Meehl (1955) noted the "understandable tendency to seek a 'construct validity coefficient'" (p. 289) but felt that given the several steps from theory derivation to hypothesis formation to observation, and given the approximate and provisional nature of the validation process, it would rarely be possible to provide such a coefficient.

Current Perspectives From Philosophy of Science

The currently predominant perspective in philosophy of science is consistent with this emphasis on theory development, with its practical implications that (a) construct validity evidence is always open to criticism and reevaluation and (b) virtually every new investigation provides a new piece of evidence pertaining to construct validation. The current perspective also does suggest an evolution in how researchers should understand construct validity. The key elements of this perspective are as follows.

Current philosophy of science emphasizes constant critical evaluation and stands in contrast to earlier philosophies of science. The earlier perspectives have been described, by Bartley (1962) and others, as justificationist: Theories could be fully justified or fully disproved based on observation or empirical evidence. The classic idea that a critical experiment could falsify or disprove a theory is an example of justificationism (Duhem, 1914/1991; Lakatos, 1968). Logical positivism (Blumberg & Feigl, 1931), with its belief that theories are straightforward derivations from observed facts, is one example of justificationist philosophy of science. Under justificationism, one could imagine the validity of a theory and its accompanying measures being fully and unequivocally established as a result of a series of critical experiments.

However, over the past 50 or 60 years, justificationism has largely been scrapped, due to advances in both philosophical work and in historical studies of how science operates (Weimer, 1979). Although there are many contentious issues in current philosophy of science (cf. Hacking, 1999; Kusch, 2002; Latour, 1999), there does appear to be general endorsement of various versions of nonjustificationism (Bartley, 1987; Campbell, 1987, 1990; Feyerabend, 1970; Kuhn, 1970; Lakatos, 1968; Weimer, 1979).

There are many aspects to nonjustificationism that pertain to construct validation (Rorer & Widiger, 1983). I will highlight one set of implications. Philosophers and historians of science recognize that the test of any theory presupposes the validity of several other theories (often referred to as auxiliary theories), including theories of measurement, that also influence the empirical test (Lakatos, 1999; Meehl, 1978, 1990a). One implication of this recognition is that a negative empirical result could reflect the failure of any number of theories other than the core proposition that led to the empirical test.

In part for this reason, no theory is ever fully proved or disproved. At any given time, evidence tends to favor some theories,

or research programs, over others. Confirming evidence can be evaluated in terms of how critical a theory test was (Meehl, 1978), and disconfirming evidence can be evaluated in terms of whether it most likely results from problems in the core theory under consideration, one of the auxiliary theories invoked to conduct the test, or another, more specific auxiliary hypothesis (Lakatos, 1968, 1999; Meehl, 1990a). Philosophers have proposed various means for evaluating the evidence (such as Lakatos's, 1999, notion of progressing vs. degenerating research programs and Meehl's, 1990a, corroboration index). Crucial to such proposals is the idea that each component of a research program, or each component of theory derivation, hypothesis formation, and empirical test, must be open to criticism. Weimer (1979) has attempted to integrate these perspectives by arguing that what characterizes science is "comprehensively critical rationalism" (p. 40), which includes the idea that every aspect of the research enterprise must be open to criticism and potential revision. As part of that process, scientists seek, strenuously, to criticize and falsify theories, while others seek, just as strenuously, to defend and verify them. In the end, each proposition and each piece of theoretical evidence is part of an argument for one theory or against another (Weimer, 1979). What makes the effort science, rather than opinion debate, is that scientists embrace critical evaluation, both in the form of theoretical argument and empirical test. And, because one can almost always defend one's theory by arguing that an apparent disconfirmation reflected a problem with an auxiliary theory or hypothesis (such as measurement), the process of theory evaluation is ongoing.

The original version of construct validity (Cronbach & Meehl, 1955), although noteworthy in its appreciation of the elusiveness of psychological constructs and the uncertainty of theory building, was more heavily influenced by justificationism than now seems warranted. The classic notion of a "nomological network" refers to lawful relations among entities and the need to place any construct in terms of its lawful relations to other constructs. The idea that we can specify a lawful network of relations and confirm nomologicals appears to imply that empirical investigations provide more certainty than we now recognize to be the case. Indeed, decades later, Meehl (1990a) referred to his earlier overemphasis on justificationism. It seems that the notion of construct validity has evolved since 1955: There is now a greater appreciation for the indeterminate, ongoing nature of theory building, theory revision, and scientific criticism.

Informative Tests of Psychological Theories

A central concern for Cronbach and Meehl (1955) was that theories concerning inferred constructs be tested with rigor. Rigor certainly referred to soundness of method, design, and test construction, but it also referred to the quality of the hypotheses one tests about a theory (Meehl, 1978, 1990a). The quality of hypothesis tests is a function of whether they facilitate the ongoing process of critical evaluation that is the hallmark of science (Weimer, 1979). To what degree does a hypothesis involve direct criticism of a theory, or direct comparison between two, alternative theoretical explanations? To what degree does a hypothesis involve a direct response to a criticism of one's theory? To what degree does a hypothesis involve a claim that, if supported, would undermine criticism of one's theory? To conduct theory tests of these kinds is to embrace the critical process that leads to advances in the truth content of psychological theories. Theory tests of this kind can be described as informative tests of theories.

One characteristic of informative theory tests is that they evaluate, as directly as possible, specific claims made for a theory. Such tests remove as many competing explanations as possible. Positive results from such tests undermine theory criticism. In clinical risk factor research, to assert that trait A is a risk factor for syndrome B requires as direct a test of that specific claim as possible. Demonstration of a positive cross-sectional correlation between A and B is not very informative. Such a test does provide information (the absence of a correlation would pose serious problems for one's risk theory), but there are many avenues open for criticism of a risk factor hypothesis. In contrast, prospective designs in which trait A predicts onset of syndrome B, and other possible explanations for the onset of B have been controlled for, are more informative. There are fewer avenues open for criticism. Tests of mediation that evaluate the putative cause, changes in the putative mediator, and changes in the putative consequence at three different sequential time points (e.g., Stice, 2001) are informative because they have taken on and ruled out some potential criticisms.

Another informative means of assessing theory claims directly is use of tests comparing alternate theoretical explanations of the same data (e.g., Bartusch, Lynam, Moffitt, & Silva, 1997). Such designs hold multiple theories up to critical examination, both individually and in comparison to each other. Doing so is an effective way to provide information to researchers and clinicians.

Both direct criticism of theories and direct responses to criticisms can be informative. Consider the psychological theory of self-enhancement, that is, the tendency to dwell on positive information about the self, rather than thoughts about one's weaknesses. The self-enhancement motive has been thought to be universal (Sedikides, Gaertner, & Toguchi, 2003). One apparently important advance from cross-cultural psychology has been the finding that, contrary to existing theory, it is not. In a series of critical tests, summarized by Heine, Lehman, Markus, and Kitayama (1999), members of collectivist cultures (such as those in eastern Asia) did not appear to self-enhance. Those tests were informative, both because they challenged the universality hypothesis directly and because outcomes of those studies could have supported universality. In a direct and critical response to that work, Sedikides et al. (2003) found evidence suggesting that members of both individualist and collectivist cultures did in fact tend to self-enhance, but they did so with respect to different behaviors. Members of individualistic cultures tended to self-enhance with respect to engaging in individualist behaviors (e.g., seeing oneself as better than others at "trust[ing] your own instinct rather than the group's instinct"), and members of collectivist cultures tended to self-enhance with respect to engaging in collectivist behaviors (e.g., seeing oneself as better than others at "defend[ing] the group's decisions"). The Sedikides et al. (2003) findings were informative because they suggested that an auxiliary hypothesis (one always self-enhances on dimensions valued by individualism) was in error, thus obscuring the true universality of self-enhancement.

In another critical turn, Heine (in press) appears to have shown that the Sedikides et al. (2003) results do not reflect self-enhancement, but rather are an artifact of their use of the "better-

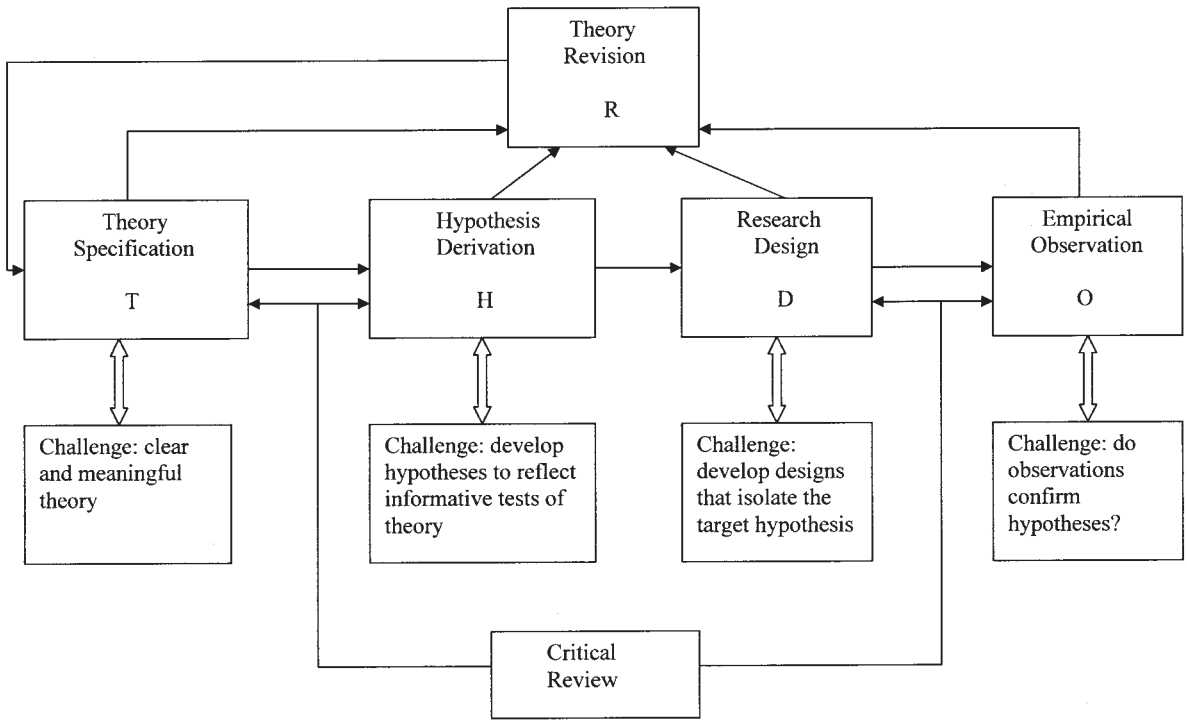

Figure 1. Depiction of the five general steps in establishing construct validity. T refers to theory, H refers to hypotheses, D refers to research design, O refers to empirical observations, and R refers to theory revisions. The figure also depicts critical review of all steps in the process. Narrow arrows refer to paths of influence; broad arrows connect a step with a statement of the challenge a researcher faces at that step.

than-average" method. The method is biased: Individuals rate themselves as better than average, they rate any random individual as better than average, and they even rate a randomly chosen fragrance as better than the average fragrance (reviewed in Heine, in press). Heine showed that members of collectivist cultures only appear to self-enhance when researchers use the better-than-average methodology, as did Sedikides et al. (2003). Thus, Heine argues, mistaken trust in the validity of methods eliciting better-than-average statements has led to the mistaken view that self-enhancement is universal. This series of tests has been quite informative.[1] Weimer's (1979) comprehensively critical rationalism describes this exchange: The core theory of self-enhancement, an auxiliary theory of the content of self-enhancement, and auxiliary theories of method were all relevant to the resolution of the issue.

Clinical assessment research has progressed toward more informative tests of clinical hypotheses. This progression has been facilitated, in part, by the acceptance of theoretical tests of inferred constructs that followed Cronbach and Meehl (1955).

A Five-Step Model for Construct Validation

To summarize, I offer a five-step model for construct validity research (depicted in Figure 1 and heavily influenced by Meehl, 1978, 1990a). The steps are (1) careful specification of the theoretical constructs in question, (2) articulation of how the theory of the construct is translated into informative hypotheses, (3) specification of appropriate research designs to test one's hypotheses, (4) articulation of how observations from samples pertain to one's predictions, and (5) revision of the theory and the constructs. The comprehensive criticism characteristic of science affects all steps of the process. Several implications of this model are apparent from the foregoing discussion. First, careful specification of theoretical constructs is crucial for clinical assessment. Clinical measures likely to make an impact are those that stem from new, clarifying, or otherwise informative theory. Second, construct validation requires informative tests, which are tests that facilitate the critical review process characteristic of science. Third, use of sound and appropriate research designs is, of course, essential for construct validation.

[1] Meehl's (1978, 1990a, 1990b, 1990c) promotion of risky tests of theories was advocacy for one type of informative test. If one hypothesized that two variables were correlated .60 or that under given risk conditions, a person's level of anxiety would be two standard deviations above the mean, one is taking a far greater risk than if one had merely hypothesized that two variables were positively related. Results of such tests are more informative: If the outcome is close to predictions, one has demonstrated much stronger support for one's theory than if one had merely confirmed a positive relationship between two variables. Although risky tests of this kind represent an important ideal for clinical research, as a practical matter they tend not to be feasible. The origins of human behavior are inherently multivariate and interactive, and they often involve dispositions that cannot be manipulated. Therefore, clinical researchers cannot exert the level of experimental control over human experience that one can exert over inanimate objects, thus reducing the level of precision of our hypotheses.

Fourth, the ability to determine how well observations from data conform to hypotheses (Step 4) is essential. Below, I provide a critical discussion of recent developments in statistical indices pertaining to hypothesis validation. I critically evaluate efforts to analyze MTMM designs, a recent suggestion by Westen and Rosenthal (2003) to quantify a measure of construct validity, and generalizability theory (GT).

Fifth, it is important to appreciate that the construct validation process involves an ongoing, iterative process in which new findings and new theories clarify and alter existing theories, thus requiring new measures and new theory tests (Haynes, Richard, & Kubany, 1995; Weimer, 1979). Ongoing revisions of theories, and the measures used to represent them, are part of the process of increasing the "truth content" (Lakatos, 1968) of clinical theories. The revision process can be triggered at any step in the construct validation process.

Philosophy of Science and the Construct Validity of Clinical Measures: Integration and Examples

In this section, I provide two kinds of examples of ways in which the concept of construct validity has benefited clinical assessment. The first concerns advances in clinical assessment theory development. The second concerns the practical process of the critical evaluation of theories from a construct validity perspective.

Advances in Clinical Assessment Due to Theory Development

With the increasing attention to theoretical development over the 50 years since Cronbach and Meehl (1955), there appears to have been a process of increasing hierarchical organization of theoretical constructs, along with progressive differentiation among lower level facets of broad constructs. Brief consideration of three examples illustrates this process. Research on self-reported mood has been enhanced by theory that hierarchically organizes mood states. Recognition of the two broad, distinct dimensions of positive affect (PA) and negative affect (NA) helped researchers to develop more precise differentiations between two lower order concepts whose measures had been confounded: anxiety and depression (Clark & Watson, 1991; Diener, Larsen, Levine, & Emmons, 1985; Watson, Clark, & Carey, 1988). The two share high, overall NA. Individuals high in NA but unremarkable in PA tend to be anxious, and those high in NA and also low in PA tend to be depressed. This work includes identification of a hierarchical, tripartite structure for distress, which includes an overall affective distress factor (high NA) and differentiation among the two lower-level, specific facets of anxiety and depression (Clark & Watson, 1991).

The organization of personality theory into comprehensive models, such as the five-factor model (Goldberg, 1990), has facilitated clinical theory development. One result has been models of personality disorders as extreme variants of combinations of personality dimensions (Widiger, Costa, & McCrae, 2002). That work, along with work using other, organized personality models (Clark, 1993; Morey et al., 2003), has helped to clarify empirical differences among correlated personality disorders, while embodying an integration between the normal and the abnormal.

Hierarchical organization and accompanying differentiation have been used recently to help explain the comorbidity of some disorders. Krueger et al. (2002) provide evidence for a common etiological contribution to externalizing disorders (such as substance abuse and antisocial behavior), along with disorder-specific etiological factors. Their work is an example of what are now common theoretical developments that seek neither to collapse syndromes into a small number of broad categories nor to insist on unique, separate disorders with their own causes. Instead, both similarities and differences between disorders are integrated into a common theory: There are common factors (and, perhaps, common causes) to sets of disorders, and there are specific factors (and, perhaps, specific causes) that differentiate disorders within a set in meaningful ways.

In sum, Cronbach and Meehl's (1955) recognition of the centrality of construct validity, and hence theory testing, in psychological inquiry has helped facilitate the development of informative, integrative clinical theory. One result has been more clear distinctions among related but separate constructs and hence more precise assessment.

The Critical Evaluation of Theories From a Construct Validity Perspective

Clinical assessment tools are constantly undergoing critical evaluation, with respect both to theoretical concerns and to the quality of supportive, empirical evidence. To illustrate the operation of this process, I briefly discuss three examples from the recent history of clinical assessment.

Faust, Hart, and Guilmette (1988) criticized neuropsychological assessment methods for failing to recognize the possibility that clients may fake head injury without detection, an obvious threat to assessment validity. This criticism is perhaps best understood as pertaining to what at the time was an important neuropsychology auxiliary hypothesis, that faking responses to neuropsychological tests would be transparent to examiners. Findings reported by Faust, Hart, and Guilmette (1988) and Faust, Hart, Guilmette, and Arkes (1988) cast serious doubt on that auxiliary hypothesis by indicating that the vast majority of neuropsychologists could not identify a faked protocol via blind interpretation, instead interpreting faked protocols as valid. In response, researchers began to develop new measures to detect faking (Hiscock & Hiscock, 1989), and now many such measures are under investigation, with encouraging results (Vickery, Berry, Inman, Harris, & Orey, 2000). As neuropsychologists discard the auxiliary hypothesis and instead develop valid means to identify faking, the validity of neuropsychological measures is enhanced, particularly in cases where test takers may be motivated by factors such as large civil judgments. This process of improvement in validity began with critical evaluation of an important auxiliary hypothesis.

The fascinating, long-lasting debate over the construct validity of the Rorschach test is a telling example of this process (Wood, Nezworski, Lilienfeld, & Garb, 2003). A test based on the idea that one's perceptions of stimuli reveal aspects of one's personality (Rorschach, 1964) and that persons project aspects of themselves onto ambiguous stimuli (Frank, 1939) was appealing, particularly in light of the failures of objective personality testing in the early part of the 20th century (Wood et al., 2003). Enthusiasm for the test became so great that it was referred to as an x-ray of the mind

(Klopfer, 1940). Unfortunately, validity studies indicated the failure of many inferences thought to follow from Rorschach responses (Cronbach, 1956; Zubin, 1954). The apparent failure of the test might have reflected the failure of the core projective hypothesis theory, failures in auxiliary theories concerning the specific nature of personality projection, or psychometric failures in operationalizing constructs and testing covariances. As a result of the negative findings, the Rorschach became steadily less prominent in the middle of the 20th century (Wood et al., 2003).

Exner's (1974, 1978) publication of a comprehensive system for the Rorschach appeared to offer sound psychometrics and good evidence for the construct validity of the test, and this led to a resurgence of interest in the Rorschach (Wood et al., 2003). Were Exner correct, the past problems with the Rorschach likely did not concern the core theory of measuring projections and personality-based perceptions, but rather the auxiliary matter of capturing its performance psychometrically. And, indeed, the Exner system received high praise for many years (Board of Professional Affairs, 1998; Butcher & Rouse, 1996).

Recently, however, the validity of Exner's system has come under serious criticism (see reviews by Hunsley & Bailey, 1999; Wood, Garb, Lilienfeld, & Nezworski, 2002; Wood et al., 2003). Many of those criticisms are psychometric, but that psychometric limitations are again in evidence raises questions about the core theory of projective testing. If repeated attempts to capture a theoretical construct psychometrically fail, the focus of criticism appropriately turns toward the theory.

Current defenses against that criticism appear to be somewhat post hoc, such as the argument that the Rorschach measures implicit personality in contrast to objective tests' focus on the explicit (Bornstein, 2001). To support this new claim, one must show that the presumably implicit measures taken from projective tests add valid information beyond what is obtained from explicit measures. Few studies have undertaken this task (although Spangler, 1992, reviewed evidence suggesting differential prediction between TAT-based and explicit aspects of the need for achievement). Overall, and certainly with respect to the Rorschach, the need for this kind of incremental validity evidence has not yet been fully met (Lilienfeld, Wood, & Garb, 2000). This current reliance on post hoc defenses reminds one of Lakatos's (1968) description of a degenerating research program (one characterized by defenses that involve a new, post hoc theoretical shift, and one unlikely to yield new knowledge or understanding). However, the debate is ongoing (Meyer, 2001). Perhaps it will lead to more valid projective tests, or perhaps it will lead to the conclusion to focus efforts elsewhere.

A third example concerns efforts to improve the validity of objective personality tests. Buss and Craik (1980, 1983) argued that trait assessments are, essentially, summary statements concerning the frequency of prototypical acts that a person engages in over time (the act frequency approach). They felt that by identifying prototypic acts for given traits, and by measuring their frequency, researchers might gain more reliable and valid assessment of traits. Initially, this idea was quite popular and received considerable attention, even gaining prominence in personality textbooks (cf. Peterson, 1988). However, Block (1989) provided a critique of all four basic construct validity steps in the act frequency validation approach. His criticisms concerned the basic theoretical approach to personality assessment (e.g., many of the "act statements" have no conceptual connection to the disposition they measure). They also pertained to the nature of the hypotheses tested and the research design (one does not study acts, but rather retrospective reports of acts) and finally, to the degree to which empirical observations supported the hypotheses (many "act statements" relate as strongly to dispositions other than those they are thought to represent as to dispositions they are thought to represent). The result has been a marked drop-off in enthusiasm for the approach, only marginally supportive evidence from its advocates (Gosling, John, Craik, & Robins, 1998), and hence, a focus of research efforts elsewhere.

As these examples illustrate, clinical assessment tools routinely undergo critical evaluation. The evaluation process simultaneously concerns core theories, auxiliary theories, and issues of method. Ultimately, the construct validity of a clinical assessment tool reflects validity on all of these levels. To date, there is no way to quantify the soundness of theory development, the validity of auxiliary theories, and the informativeness of theory tests (see Campbell, 1990; Fiske, 1990; Meehl, 1990a, 1990b; Serlin & Lapsley, 1990, for discussion of quantification efforts). Researchers do, however, routinely apply statistical analyses to quantify the degree to which observations conform to predictions (Step 4 in the construct validity process). In the next section, I consider ongoing efforts to improve researchers' ability to do so. I focus on important advances in this domain that can enhance the construct validation process.

Fit Between Observations and Hypotheses: Statistical Approaches to the Fourth Stage of Construct Validation

One of the primary means by which researchers have sought to quantify Step 4 evidence has been by developing statistical means for evaluating MTMMs (Cudeck, 1988; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Marsh & Grayson, 1995; Reichardt & Coleman, 1995). Recently, a different approach has been proposed by Westen and Rosenthal (2003), which involves calculating simple correlation indices of construct validity, or more precisely, indices of fit between hypotheses and observations. GT also merits consideration in this context (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989), because applications of GT enable researchers to identify and quantify multiple sources of variance in scores. In this section, I briefly describe current MTMM approaches, Westen and Rosenthal's (2003) construct validity statistic, and contributions from GT. I then critically evaluate the impact of each on the fourth phase of construct validation.

Procedures for Analyzing MTMMs

Campbell and Fiske's (1959) classic description of the MTMM design was a profoundly important methodological suggestion for improving the investigation of fit between hypotheses and observations. Following Cronbach and Meehl's (1955) discussion of the importance of both convergent and discriminant validity to isolate the meaning of a construct measure, Campbell and Fiske (1959) highlighted a serious threat to Step 4 analyses: the extent to which reliable variance on a measure is due to the method of assessment, rather than to the targeted construct. They noted that convergent validity coefficients often failed to exceed correlations between

two presumably unrelated traits that shared only a common assessment method, thus illustrating the sizable contribution of method variance to overall measure variance. Thus, statistically significant convergent validity correlations may overestimate true effects. (It should be noted that Campbell and Fiske's, 1959, design is not limited to the analysis of traits. Any set of constructs measured in multiple ways can be examined with the MTMM design. In the discussion that follows, *clinical attribute* can be substituted for *trait*.) The MTMM design had elegance, but of course, the means of analyzing its results was informal. They relied heavily on visual inspection of correlation coefficients.

Today, with ready access to high-speed computers that can run statistical software of enormous complexity, investigators have developed relatively straightforward means of evaluating the goodness of fit of complex models. A number of well-performing fit indices have been developed for use with structural equation modeling (SEM; Bentler & Wu, 1995); those indices mark how well a pattern of obtained covariances fit the predicted covariances—thereby facilitating Step 4 evaluation. There has been an accompanying explosion of statistical developments for evaluating MTMMs (Shrout & Fiske, 1995). One of the first such approaches was to test models holding that responses to any item can be understood as reflecting additive effects of trait variance, method variance, and measurement error (Marsh & Grayson, 1995; Reichardt & Coleman, 1995; Widaman, 1985). In this approach, both trait and method factors are modeled explicitly. Thus, if indicator X reflects method A for evaluating trait A, that part of the variance of X that is shared with other indicators of trait A is assigned to a trait A factor, that part of the variance of X that is shared with indicators of other constructs measured by method A is assigned to a method A factor, and the remainder is assigned to an error term (Eid et al., 2003; Kenny & Kashy, 1992). The association of each type of factor with other measures can be examined, so, for example, one can model explicitly the role of a certain trait or a certain type of method variance on responses to a criterion measure. Other approaches recognize interactions between traits and methods (Campbell & O'Connell, 1967, 1982) and therefore test multiplicative models (Browne, 1984; Cudeck, 1988).

The theoretical advantages of this type of approach for Step 4 evaluation are clear. Relying only on trait variance, one could evaluate the overall fit of data to predictions based on a model. Overall fit indexes help identify broad discrepancies between hypotheses and observations. One could also examine each covariance individually to identify which specific relationships within the model did not conform to predictions, using statistical significance criteria. Thus, one could formally identify which specific findings are inconsistent with Step 4 hypothesis evaluation.

As exciting as this prospect is, it has generally turned out not to be feasible. As Kenny and Kashy (1992) describe, in this approach one models more factors than there is information to identify them (referred to as overfactoring). One therefore often finds either impossible values (such as negative variances) or a failure of one's computer program to converge on a solution (Kenny, 1995). As a result, an alternative approach has become popular: Instead of modeling method factors, one identifies the presence of method variance by determining whether the residual variances of construct indicators that share the same method are correlated, after accounting for construct variation and covariation. If so, method variance has been captured in the model (Marsh & Grayson, 1995). This "correlated uniquenesses" approach models only the trait factors, so it avoids the overfactoring problem referred to above. On the other hand, there is an important limitation to this approach. Without method factors, one cannot examine the association of method variance with other constructs, which may be important to do (Cronbach, 1995).

Most recently, Eid et al. (2003) have offered a new, alternative approach that appears to avoid overfactorization yet enables modeling of some method variance. Essentially, they suggest modeling all trait factors and all but one method factor. The practical result is there are fewer factors, and the resulting models appear to be identified. One theoretical implication is that one method is chosen as the baseline method, and one evaluates other methods for how they influence results compared to the baseline method. Suppose, for example, that one had anonymous questionnaire and clinical interview data for a series of traits. One might specify the questionnaire method as the baseline method, so a questionnaire method factor is not modeled as separate from trait variance, and trait scores are really trait-as-measured-by-questionnaire scores. One then models a method factor for clinical interview. If clinical interview leads to lower trait reporting than does the anonymous questionnaire, one would find that the interview method factor correlated negatively with the trait in question (or, the trait-as-measured-by-questionnaire score). That would imply that individuals report lower levels of a trait during an interview than they report in questionnaires. Further, one can assess whether this process works differently for different traits. Perhaps the clinical interview method lowers reports of some traits more than others. Such a possibility can be examined empirically using this method. Thus, this approach appears to hold the promise of identifying the contribution of method to measure scores, although it has the limitation that the choice of "baseline method" influences the results and may be arbitrary (Eid et al., 2003).

It is also the case that as useful as MTMM designs are for Step 4 construct validity analyses, they may lack clinical meaning. By itself, the design does not include differential prediction of clinical outcome by the different traits. For clinical assessment, the value of assessing clinically relevant attributes is often that they enable prediction of some criterion of clinical importance. Hammond, Hamm, and Grassia (1986) offered an approach for combining the convergent and discriminant validity of the MTMM design with evaluation of differential prediction of outcomes of interest. They describe a performance validity matrix, which adds criterion variables for each trait to the MTMM design. For example, Fischer, Smith, and Cyders (2004) demonstrated the convergent and discriminant validity of questionnaire and interview means of measuring four distinct, impulsivity-like constructs: lack of planning (acting without thinking), sensation seeking (seeking new and novel stimulation), lack of perseverance (inability to sustain attention to a task), and urgency (acting rashly in response to subjective distress). They then applied Hammond et al.'s (1986) performance validity matrix concept and showed that each trait predicted different outcomes (e.g., sensation seeking uniquely predicted frequency of drinking and gambling, and urgency uniquely predicted problem drinking and problem gambling). By applying a performance validity component, one can extend the network of hypothesis tests and thus provide more extensive information about a model.

In sum, researchers are encouraged to consider the Eid et al. (2003) approach for Step 4 construct validity analyses and to include, where appropriate, performance validity evaluation as described by Hammond et al. (1986).

Westen and Rosenthal (2003): The Quantification of Construct Validity

A different type of approach to quantifying Step 4 analytic results was recently proposed by Westen and Rosenthal (2003). They expressed concern about the informal means by which researchers determine whether a measure has construct validity. As they noted, researchers often examine a set of correlations to judge, somewhat subjectively, whether those correlations are sufficiently close to theoretical predictions to justify the conclusion that a target measure appears to have construct validity. In response to the subjective, and hence vague, quality of the validity evaluation process, they advocated for quantifying construct validity. By quantifying construct validity, they meant quantifying the degree to which one accurately predicted the correlations obtained in a typical convergent-discriminant correlation matrix. In essence, they argued for quantification of Step 4 in the construct validation process. Their interest was not an entire MTMM but rather the construct validity of a single measure.

They argued for the use of two simple correlation coefficients. The first, labeled $r_{\text{alerting-cv}}$, is computed as follows. One specifies a predicted set of convergent and discriminant correlations and then correlates that set of predicted values with the obtained values (using appropriate weights and *r*-to-*z* transformations). They call it an "alerting correlation" because it is a "rough, readily interpretable index that can alert the researcher to possible trends of interest" (Westen & Rosenthal, 2003, p. 610).

The second, labeled $r_{\text{contrast-cv}}$, involves contrast tests of correlations. For instance, suppose one hypothesizes one set of positive correlations between a target measure and certain variables and one set of negative correlations between the target measure and other variables. Those predicted correlations, represented as lambda weights, are multiplied by the obtained correlations, represented in *z* form. One obtains a contrast coefficient by summing those products. If one has accurately predicted which correlations are positive and which are negative, then positive lambda weights are multiplied by positive obtained correlations, negative lambda weights are multiplied by negative correlations, and the sum yields a highly positive contrast coefficient. The logic is the same as that for contrasting means in analysis of variance. To be more concrete, summing the products of the entries in Columns 3 and 4 of Table 1 ("Lambda weights" and "Obtained *z* correlations") gives the contrast coefficient Westen and Rosenthal (2003) used for calculating an example of $r_{\text{contrast-cv}}$. The statistic $r_{\text{contrast-cv}}$ is a function of the contrast coefficient, the intercorrelations among the variables, and the absolute values of the correlations between the target measure and its criteria. Just as with an analysis of variance contrast, it is influenced by sample size (Westen & Rosenthal, 2003).

Westen and Rosenthal (2003) offer one example of calculating the two correlations using adolescent personality disorder data (Westen, Shedler, Durrett, Glass, & Martens, 2003). They studied a new personality disorder diagnosis, "histrionic personality disorder of adolescence," by relating it to 10 existing adult personality disorder diagnoses. The 10 were chosen to reflect either convergent or discriminant validity. Using ratings to reflect each disorder, they found an $r_{\text{alerting-cv}}$ of .90 and an $r_{\text{contrast-cv}}$ of .72. Clearly, their quantification of Step 4 construct validity evidence yielded impressively high values. They concluded that "the magnitude and

Table 1
Construct Validity Analysis Results From Westen and Rosenthal (2003) and From Three Possible Alternative Sets of Findings

Diagnoses involved in prediction	Westen and Rosenthal's predicted correlations	Lambda weights	Obtained *z* correlations	Alternative correlations		
				1	2	3
Histrionic	.60	7	.62	.62	.00	.30
Borderline	.30	4	.56	.00	.00	.30
Dependent	.10	2	.20	.00	.00	.30
Antisocial	.00	1	−.06	.00	.00	.30
Narcissistic	.00	1	.10	.00	.00	.30
Paranoid	−.10	0	−.04	.00	−.04	−.30
Obsessive–compulsive	−.40	−3	−.23	.00	−.23	−.30
Avoidant	−.50	−4	−.20	.00	−.20	−.30
Schizoid	−.50	−4	−.15	.00	−.15	−.30
Schizotypal	−.50	−4	−.02	.00	−.02	−.30
$r_{\text{alerting-cv}}$			.90	.65	.69	.84
$r_{\text{contrast-cv}}$			.72*	.39*	.20*	.70*

Note. Columns 1–4 are reproduced with permission from "Quantifying Construct Validity: Two Simple Measures," by D. Westen and R. Rosenthal, 2003, *Journal of Personality and Social Psychology, 84,* pp. 612–613. Alternative correlations are hypothetical, alternative values to those reported by Westen and Rosenthal (2003). The numbers are *z* transformations of correlations.
* $p < .001$.

meaning of these *r*s . . . suggest that we understood the construct very well" (Westen & Rosenthal, 2003, p. 612).

To appreciate accurately the implications of these correlations, one must have a clear understanding of the meaning of the two coefficients. It appears that what Westen and Rosenthal (2003) meant by quantifying the degree of accuracy of prediction was accuracy in predicting the relative magnitude of the observed correlations. They noted that researchers are seldom in a position to predict precise magnitude of correlations with great accuracy.[2] And, indeed, $r_{\text{alerting-cv}}$ reflects the magnitude of the correlations only in the sense that it responds to their relative magnitude: Consistent relative magnitude of predicted and obtained correlations will produce high correlations, regardless of absolute magnitude. $r_{\text{contrast-cv}}$ is sensitive to the overall magnitude of the contrasted correlations (Westen & Rosenthal, 2003) but is not an index of individual departures from predicted magnitudes. In addition, because both indices quantify predictive accuracy with a single number, their method does not aid in the formal identification of which correlations fail to support Step 4 construct validity.

The idea behind these indices is nevertheless appealing: One must commit oneself to specific hypotheses about predicted relationships, and one gets a formal measure of success. Unfortunately, there are difficulties with their correlations, so that they can produce overly optimistic estimates of Step 4 success.

Westen and Rosenthal (2003) reported excellent results from their construct validity analysis of adolescent histrionic personality disorder: one correlation of .90 and another of .72 (using ratings, not number of symptoms, to measure the disorders; the correlations were higher when number of symptoms was used). Correlations that high are rare in psychology. However, they did not provide any other examples to show the likely range of $r_{\text{alerting-cv}}$ and $r_{\text{contrast-cv}}$ for other possible study outcomes. Table 1 helps address that need. The first four columns of the table are reproduced from Westen and Rosenthal (2003). They present the diagnoses that the target measure was correlated with, the predicted correlations, the lambda weights that reflect those correlations, and the obtained correlations (transformed into *z* scores). Below the obtained correlations are listed the values of $r_{\text{alerting-cv}}$ and $r_{\text{contrast-cv}}$.

Columns 5, 6, and 7 present three other possible outcomes. For these hypothetical examples, I have followed the example cited above by using their coefficients based on ratings to measure the disorders, and I have presumed the same average intercorrelation among the predictor variables as that reported by Westen and Rosenthal (2003; $r = .113$). In Column 5, the hypothetical situation depicted is one in which the adolescent histrionic measure correlated only with the adult histrionic measure and, counter to predictions, did not correlate with any of the other nine adult personality disorder scores. In this imagined case, only one convergent validity correlation fit predictions, and two discriminant validity correlations fit (both predicted to be 0). The correlation used for the histrionic criterion score is that reported by Westen and Rosenthal (2003) and is again presented in *z* score form. As the table indicates, $r_{\text{alerting-cv}}$ for this case was .65. Thus, in a situation in which the only significant correlation with the new adolescent measure was with the adult measure it was based on, a correlation designed to quantify Step 4 construct validity evidence appears to be quite substantial. In this hypothetical case, the $r_{\text{contrast-cv}}$ was .39, which is statistically significant ($p = 3.37\text{e-}11$).

In Column 6, an imaginary pattern is depicted in which the adolescent histrionic measure correlated .00 with all of the criterion variables, thus failing to conform to convergent validity predictions. However, the measure did correlate negatively with each of five measures, as hypothesized (again, using the correlations presented in Westen & Rosenthal, 2003). In that case, $r_{\text{alerting-cv}} = .69$, and $r_{\text{contrast-cv}} = .20$, $p = .0005$). One could argue that high values are appropriate here, as it appears that 7 of the 10 predictions were borne out (including the 2 predicted to be .00). On the other hand, there is no convergent validity in this example, even between the adolescent and adult versions of the same measure. Such a pattern does not reflect good Step 4 construct validity evidence, even though it does reflect good discriminant validity.

In Column 7, a situation is depicted in which hypotheses about direction of relationship were reasonably well borne out, but the adolescent histrionic measure had the same magnitude of relation with all 10 adult measures. In this hypothetical case, the evidence is inconsistent with the notion that one has uniquely measured adolescent histrionic personality disorder, yet values of both $r_{\text{alerting-cv}}$ and $r_{\text{contrast-cv}}$ were virtually as high as those in Westen and Rosenthal's (2003) example ($r_{\text{contrast-cv}}$ was significant at $p = 5.18\text{e-}41$).

One can see why one gets such high correlations in cases where there is little agreement between predictions and observations. The $r_{\text{alerting-cv}}$ statistic is based on a very small sample size: In these examples, it is calculated on only 10 associations. In practice, 10 probably represents the upper end in terms of the number of correlations typically reported in studies describing Step 4 construct validity findings. With such a small sample size, one can easily have, as constructed in Table 1, Column 5, a case in which there appears to be no relationship at all when considering 9 of the 10 associations, but when the 10th is included, the overall relationship appears quite strong. In many situations one would consider the resulting high correlation spurious or, more cautiously, unconvincing. The $r_{\text{contrast-cv}}$ statistic essentially contrasts two sets of correlations: If, as in the Column 5 example, all correlations in one set are 0, and only one in the contrasting set is very different from 0, a statistical contrast does in fact exist. High construct validity correlations do not necessarily reflect patterns of associations consistent with convincing evidence of Step 4 construct validation.

Although Westen and Rosenthal (2003) argued that researchers should consider seriously even nonsignificant or small-seeming $r_{\text{contrast-cv}}$ values as important, the hypothetical examples provided

[2] Westen and Rosenthal (2003) suggested one might measure the distance, D, between predicted and obtained *z* values to assess prediction of correlation magnitude. Reviewer William Grove offered the similar suggestion that one might consider the squared difference between predicted and observed correlations. As that value increases, agreement between hypotheses and observations drops. In a similar vein, an anonymous reviewer suggested the use of multiple intraclass correlations, because different versions of intraclass correlations are differentially sensitive to rank-order differences, differences in magnitude, and even differences in predicted and observed degrees of variance in measures (McGraw & Wong, 1996). Developing a magnitude-sensitive measure of some kind would, of course, create a more exacting standard for comparing hypotheses and observations.

here have produced highly significant $r_{contrast\text{-}cv}$ values even without good evidence of successful Step 4 construct validation. In practice, one would of course examine the pattern of correlations qualitatively for their substantive meaning (Westen & Rosenthal, 2003), which would prevent one from interpreting my hypothetical examples as indicative of good construct validity. But since one would do so, and since high $r_{alerting\text{-}cv}$ and $r_{contrast\text{-}cv}$ values do not necessarily reflect good construct validity, calculating those statistics may not greatly influence one's inferences regarding construct validity.

GT and Clinical Assessment

GT can have an important impact on one's ability to compare hypotheses to observations. The basic GT notion is to design studies and conduct statistical analyses so that one can isolate and quantify test response variability due to each of several factors, such as the person, the items, the occasion, the interviewer, the raters of interviews, and so on. One can vary each of those factors in one design and use analysis of variance to estimate the degree of test score variability due to each of those factors and the interactions among them (Cronbach et al., 1972).

One impact of GT on statistical estimates of validity concerns GT's relation to the classic test theory concept of reliability. The core logic is that when one wants to know whether scores are reliable, one's basic concern is whether scores generalize across some dimension, be it items (internal consistency reliability), occasions (test–retest reliability), interviewers, raters, or other factors. By estimating the different sources of variance in test scores, one can make comparisons, such as comparing the variability due to individual differences to the variability due to raters. In that way, one can determine the generalizability (reliability) of scores across whichever dimension is of interest. The notion of generalizability thus reflects a broader concept than that of reliability: Any one reliability analysis concerns one specific form of generalizability (Cronbach et al., 1972; Shavelson et al., 1989).

There are numerous advantages to this approach. By quantifying the influence of each of several factors on individual differences in responses, one has more comprehensive measurement information than one has from calculating one, or even two, reliability estimates. As a result, one can identify which sources of variance it is most important to attend to in future studies. For example, if there is significant variability across raters, then averaging the responses of multiple raters will significantly increase the reliability of ratings. In addition, two different research designs using the same measure might involve different sources of measurement error: After a generalizability study, one has an estimate of the degree of error each source brings. Thus, applications of GT give one a much greater capacity to control important measurement error than one has after simply calculating one reliability estimate, such as internal consistency.

The role of GT in clinical assessment research is difficult to determine. Few clinical assessment studies report GT-based findings. Perhaps, after classic reliability analyses yield solid evidence of internal consistency and stability over time, many researchers judge that the further information provided by a generalizability study does not warrant the necessary allocation of resources. Perhaps researchers prefer to proceed to tests of substantive validity hypotheses instead. However, there are contexts in which GT is uniquely helpful, and those are the settings in which GT does tend to be applied. Investigations that must consider individual difference variance along with multiple other sources of variance, such as items and raters (Trusty, Burger, Calsyn, Klinkenberg, & Morse, 1996); situations, response classes, and types of data (Vandambaggen, Vanheck, & Kraaimaat, 1992); multiple raters in different situations (Gerlsma, Snijders, vanDuijn, & Emmelkamp, 1997; Lavigueur, Tremblay, & Saucier, 1993); and gender of rater and subject (Davidson et al., 1996), tend to be those that conduct generalizability studies. When one does need to consider multiple sources of variance, investment in a generalizability study is worthwhile, because one can obtain estimates of all relevant influences.

There is another way in which the core GT notion of identifying and controlling many influences on test responses has become typical in clinical assessment research. The capacity researchers now have, using SEM, to study the influence of only the shared variance among indicators of a construct enables them to eliminate both random error and systematic variance unique to an indicator (that therefore reflects some other variance source). In fact, the application of SEM to MTMM designs (Eid et al., 2003) enables one to go further than what was possible when GT was developed. One can estimate multiple sources of method variance on a construct indicator and then estimate the influence of those sources of variance on other factors. For instance, as noted above, one can estimate not only variance due to use of an interview method but also differences in the influence of the interview method on the assessment of different clinical attributes (Eid et al., 2003).

Similarly, the use of individual growth curve models for longitudinal data enables one to model a general, change-over-time factor and then consider individual difference factors that cause variability around the average change (Duncan, Duncan, Strycker, & Li, 1999). This modeling can be done within an SEM framework, allowing one to simultaneously estimate variability due to occasion, due to items, and due to the target individual differences. Thus, although classic GT studies are the exception, not the norm, the process of systematically modeling and investigating multiple influences on a test response (the heart of GT) is becoming the norm, not the exception. Doing so can greatly improve the accuracy of researchers' estimates of concordance between hypotheses and observations.

In sum, although not all attempts to measure Step 4 construct validity success have been successful, clear progress is being made. Clinical researchers can isolate different sources of variance in test responses and examine the influence of those variance sources on other factors. Continued efforts to improve these measurements are part of the legacy of Cronbach and Meehl (1955).

Summary

Cronbach and Meehl (1955) recognized that psychologists study inferred or nonobservable constructs. They observed that for such hypothetical constructs, the only way to determine whether a measure reflects a construct validly is to test whether scores on the measure conform to a theory, of which the target construct is a part. Construct validity is thus basic theory testing in psychology. Determining whether a measure is a valid representation of a hypothetical construct is part of the process of theory testing.

In the 50 years since publication of their article, philosophy of science has evolved further in directions implied by their work. There has been a growing recognition that virtually every theory test necessarily invokes numerous auxiliary theories and specific auxiliary hypotheses. Therefore, results of any theory test may pertain to the target theory, but they may pertain instead to any number of auxiliary theories or hypotheses. As a result, theories are not fully proved or disproved. Instead, science is characterized by the ongoing, comprehensive process of critical evaluation of all phases of scientific inquiry (Weimer, 1979). The construct validity of clinical measures thus refers to an ongoing process of discovery, pertaining both to theories and the measures that embody them. One result of these developments is that clinical researchers engage in increasingly informative evaluations of theories and the measures that accompany them.

There are at least five steps in construct validity work: careful theory specification, development of informative hypothesis tests, use of sound research design, examination of the degree to which observations confirm hypotheses, and ongoing revisions of both theory and measures.

Recognition of the importance of theory has led to valuable advances in clinical assessment. Understanding of psychopathology appears to be evolving away from a state of isolated hypotheses and conceptual frameworks, toward more comprehensive, hierarchically organized explanatory frameworks. One advantage of such integrative frameworks is that they facilitate differentiation among related, lower level facets of broader constructs. Clinical prediction is often improved with such differentiation (Smith et al., 2003). The critical evaluation of clinical assessment models concerns all stages in the construct validity process: Theories, hypotheses, designs, and specific measures are all held to critical scrutiny.

Considerable attention has been paid to means of evaluating the degree to which empirical observations conform to hypotheses; researchers are seeking both more precise and more comprehensive means for conducting such analyses. Recent advances in applying SEM to MTMM designs can be combined with models for increasing the representativeness of designs to provide more accurate evaluations of the validity of clinical assessment methods (Eid et al., 2003; Hammond et al., 1986). One can use such methods to isolate different sources of variance in clinical assessment procedures and examine their influence on clinical prediction. These approaches provide one way to identify the multiple influences on test scores, which is a central goal of GT, and they go further by enabling one to examine the predictive role of those various influences. These new tools have both obvious theoretical importance (concerning the validity of measures) and clear practical importance (concerning accurate, applied assessment).

Westen and Rosenthal (2003) recently offered an attempt to quantify construct validity, by which they meant quantify the fit between observations and hypotheses (Step 4 in the current model). They made a compelling case for the need for more precise, formal evaluation of validity data. However, the specific indices they proposed can give overly optimistic estimates of fit, so high values cannot, by themselves, be interpreted as evidence of construct validity.

In sum, there are numerous advances in clinical assessment research that stem, at least in part, from the seminal work of Cronbach and Meehl (1955). Valuable theoretical advances have accrued, and researchers have begun to develop more accurate means of evaluating validity evidence. Ongoing, critical evaluation and hence evolution in assessment knowledge appears to be the norm, and even our understanding of the concept of construct validity continues to evolve. Researchers are encouraged to embrace these perspectives and thus facilitate further advances in the validity of clinical assessment.

References

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement, 10,* 67–78.

Bartley, W. W., III. (1962). *The retreat to commitment.* New York: A. A. Knopf.

Bartley, W. W., III. (1987). Philosophy of biology versus philosophy of physics. In G. Radnitzky & W. W. Bartley, III (Eds.), *Evolutionary epistemology, rationality, and the sociology of knowledge* (pp. 7–46). La Salle, IL: Open Court.

Bartusch, D. R. J., Lynam, D. R., Moffitt, T. E., & Silva, P. A. (1997). Is age important? Testing a general versus a developmental theory of antisocial behavior. *Criminology, 35,* 13–48.

Bentler, P. M., & Wu, E. J. C. (1995). *Structural equations program manual.* Los Angeles: BMDP Software.

Block, J. (1989). Critique of the act frequency approach to personality. *Journal of Personality and Social Psychology, 56,* 234–245.

Blumberg, A. E., & Feigl, H. (1931). Logical positivism. *Journal of Philosophy, 28,* 281–296.

Board of Professional Affairs. (1998). Awards for distinguished professional contributions: John Exner. *American Psychologist, 53,* 391–392.

Bornstein, R. F. (2001). The clinical utility of the Rorschach Inkblot Method: Reframing the debate. *Journal of Personality Assessment, 77,* 39–47.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology, 37,* 1–21.

Buss, D. M., & Craik, K. H. (1980). The frequency concept of disposition: Dominance and prototypically dominant acts. *Journal of Personality, 43,* 379–392.

Buss, D. M., & Craik, K. H. (1983). Act prediction and the conceptual analysis of personality scales: Indices of act density, bipolarity, and extensity. *Journal of Personality and Social Psychology, 45,* 1081–1095.

Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology, 47,* 87–111.

Campbell, D. T. (1987). Evolutionary epistemology. In G. Radnitzky & W. W. Bartley, III (Eds.), *Evolutionary epistemology, epistemology, rationality, and the sociology of knowledge* (pp. 47–89). La Salle, IL: Open Court.

Campbell, D. T. (1990). The Meehlian corroboration-verisimilitude theory of science. *Psychological Inquiry, 1,* 142–147.

Campbell, D. T. (1995). The postpositivist, nonfoundational, hermeneutic epistemology exemplified in the works of Donald W. Fiske. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 13–28). Hillsdale, NJ: Erlbaum.

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

Campbell, D. T., & O'Connell, E. J. (1967). Method factors in multitrait-multimethod matrices: Multiplicative rather than additive? *Multivariate Behavioral Research, 2,* 409–426.

Campbell, D. T., & O'Connell, E. J. (1982). Methods as diluting trait relationships rather than adding irrelevant systematic variance. In D. Brinberg & L. Kidder (Eds.), *New directions for methodology of social and behavioral science: Forms of validity in research* (pp. 93–111). San Francisco: Jossey-Bass.

Clark, L. A. (1993). *Manual for the schedule for nonadaptive and adaptive personality.* Minneapolis: University of Minnesota Press.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100,* 316–336.

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

Cronbach, L. J. (1956). Assessment of individual differences. *Annual Review of Psychology, 7,* 173–196.

Cronbach, L. J. (1985, June). *Construct validation after thirty years.* Paper presented at the University of Illinois, Department of Educational Psychology, Champaign, IL.

Cronbach, L. J. (1995). Giving method variance its due. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 145–160). Hillsdale, NJ: Erlbaum.

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

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

Cudeck, R. (1988). Multiplicative models and MTMM matrices. *Journal of Educational Statistics, 13,* 131–147.

Davidson, K., MacGregor, M. W., MacLean, D. R., McDermott, N., Farquharson, J., & Chaplin, W. F. (1996). Coder gender and potential for hostility ratings. *Health Psychology, 15,* 298–302.

Diener, E., Larsen, R. J., Levine, S., & Emmons, R. A. (1985). Intensity and frequency: Dimensions underlying positive and negative affect. *Journal of Personality and Social Psychology, 48,* 1253–1265.

Duhem, P. (1991). *The aim and structure of physical theory* (P. Weiner, Trans.). Princeton, NJ: Princeton University Press. (Original work published 1914)

Duncan, T. E., Duncan, S. C., Strycker, L. A., & Li, F. (1999). *An introduction to latent growth curve modeling.* Mahwah, NJ: Erlbaum.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8,* 38–60.

Exner, J. E. (1974). *The Rorschach: A comprehensive system* (Vol. 1). New York: Wiley.

Exner, J. E. (1978). *The Rorschach: A comprehensive system. Vol. 2: Current research and advanced interpretation.* New York: Wiley.

Faust, D., Hart, K., & Guilmette, T. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology, 56,* 578–582.

Faust, D., Hart, K., Guilmette, T., & Arkes, H. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology: Research and Practice, 19,* 508–515.

Feyerabend, P. (1970). Against method. In M. Radner & S. Winokur (Eds.), *Minnesota studies on the philosophy of science: Vol. IV. Analyses of theories and methods of physics and psychology* (pp. 17–130). Minneapolis, MN: University of Minnesota Press.

Fischer, S., Smith, G. T., & Cyders, M. A. (2004, November). Impulsivity: Construct validation of four types and implications for comorbidity of gambling, drinking, and binge eating. Paper presented at the annual meeting of the Association for the Advancement of Behavior Therapy, New Orleans, LA.

Fiske, D. W. (1990). Judging results and theories. *Psychological Inquiry, 1,* 151–152.

Fiske, D. W. (1995). Reprise, new themes, and steps forward. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 351–362). Hillsdale, NJ: Erlbaum.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7,* 248–260.

Frank, L. (1939). Projective methods for the study of personality. *Journal of Psychology, 8,* 389–413.

Gerlsma, C., Snijders, T. A. B., vanDuijn, M. A. J., & Emmelkamp, P. M. G. (1997). Parenting and psychopathology: Differences in family members' perceptions of parental rearing styles. *Personality and Individual Differences, 23,* 271–282.

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

Goldberg, L. R. (1995). Reprise, new themes, and steps forward. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 29–44). Hillsdale, NJ: Erlbaum.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line coding by observers. *Journal of Personality and Social Psychology, 74,* 1337–1349.

Guion, R. M., & Cranny, C. J. (1982). A note on concurrent and predictive validity designs: A critical re-analysis. *Journal of Applied Psychology, 67,* 239–244.

Hacking, I. (1999). *The social construction of what?* Cambridge, MA: Harvard University Press.

Hammond, K. R., Hamm, R. M., & Grassia, J. (1986). Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychological Bulletin, 100,* 257–269.

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

Heine, S. J. (2005). Where is the evidence for pan-cultural self-enhancement? A reply to Sedikides, Gaertner, and Toguchi. *Journal of Personality and Social Psychology, 89,* 531–538.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive regard? *Psychological Review, 106,* 766–794.

Hiscock, M., & Hiscock, C. (1989). Refining the forced choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11,* 967–974.

Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment, 11,* 266–277.

Kenny, D. A. (1995). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 111–124). Hillsdale, NJ: Erlbaum.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Klopfer, B. (1940). Personality aspects revealed by the Rorschach method. *Rorschach Research Exchange, 4,* 26–29.

Krueger, R. F., Hicks, B. M., Patrick, C. J., Carlson, S. R., Iacono, W. G., & McGue, M. (2002). Etiologic connections among substance dependence, antisocial behavior, and personality: Modeling the externalizing spectrum. *Journal of Abnormal Psychology, 111,* 411–424.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kusch, M. (2002). Metaphysical déjà vu: Hacking and Latour on science studies and metaphysics. *Studies in History and Philosophy of Science, 33,* 639–647.

Lakatos, I. (1968). Criticism and the methodology of scientific research programs. *Proceedings of the Aristotelian Society, 69,* 149–186.

Lakatos, I. (1999). Lectures on scientific method. In I. Lakatos & P.

Feyerabend (Eds.) *For and against method* (pp. 19–112). Chicago: University of Chicago Press.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41,* 1183–1192.

Latour, B. (1999). *Essays on the reality of science studies.* Cambridge, MA: Harvard University Press.

Lavigueur, S., Tremblay, R. E., & Saucier, J. F. (1993). Can spouse support be accurately and reliably rated? A generalizability study of families with disruptive boys. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 34,* 689–714.

Lawshe, C. L. (1985). Inferences from personnel tests and their validities. *Journal of Applied Psychology, 70,* 237–238.

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

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and application* (pp. 177–198) London: Sage.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Karl, Ronald, and slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141.

Meehl, P. E. (1990b). Author's response. *Psychological Inquiry, 1,* 173–180.

Meehl, P. E. (1990c). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244.

Meehl, P. E. (1995). Utiles, hedons, and the mind-body problem, or, who's afraid of Vilfredo? In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 45–66). Hillsdale, NJ: Erlbaum.

Messick, S. (1980). Test validity and ethics of assessment. *American Psychologist, 35,* 1012–1027.

Meyer, G. J. (2001). Evidence to correct misperceptions about Rorschach norms. *Clinical Psychology: Science and Practice, 8,* 389–396.

Morey, L. C., Warner, M. B., Shea, M. T., Gunderson, J. G., Sanislow, C. A., Grilo, C., et al. (2003). The representation of four personality disorders by the schedule for nonadaptive and adaptive personality dimensional model of personality. *Psychological Assessment, 15,* 326–332.

Peterson, C. (1988). *Personality.* New York: Harcourt Brace Jovanovich.

Reichardt, C. S., & Coleman, S. C. (1995). The criteria for convergent and discriminant validity in a multitrait-multimethod matrix. *Multivariate Behavioral Research, 30,* 513–538.

Rorer, L., & Widiger, T. (1983). Personality structure and assessment. *Annual Review of Psychology, 34,* 431–463.

Rorschach, H. (1964). *Psychodiagnostics.* New York: Grune & Stratton.

Sedikides, C., Gaertner, L., & Toguchi, T. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology, 84,* 60–79.

Serlin, R. C., & Lapsley, D. K. (1990). Meehl on theory appraisal. *Psychological Inquiry, 1,* 169–172.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44,* 922–932.

Shrout, P. E. (1995). Measuring the degree of consensus in personality judgments. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 79–92). Hillsdale, NJ: Erlbaum.

Shrout, P. E., & Fiske, S. T. (1995). *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske.* Hillsdale, NJ: Erlbaum.

Smith, G. T., Fischer, S., & Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment, 15,* 467–477.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300–308.

Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin, 112,* 140–154.

Stice, E. (2001). A prospective test of the dual-pathway model of bulimic pathology: Mediating effects of dieting and negative affect. *Journal of Abnormal Psychology, 110,* 124–135.

Trusty, M. L., Burger, G. K., Calsyn, R. J., Klinkenberg, W. D., & Morse, G. A. (1996). Generalizability of the scales of the original and expanded versions of the Brief Psychiatric Rating Scale. *International Journal of Methods in Psychiatric Research, 6,* 195–201.

Vandambaggen, R., Vanheck, G. L., & Kraaimaat, F. (1992). Consistency of social anxiety in psychiatric patients: Properties of persons, situations, response classes, and types of data. *Anxiety, Stress, and Coping, 5,* 285–300.

Vickery, C. D., Berry, D. T. R., Inman, T. H., Harris, M. J., & Orey, S. A. (2000). Detection of inadequate effort on neuropsychological testing: A meta-analytic review of selected procedures. *Archives of Clinical Neuropsychology, 16,* 45–73.

Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affect and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology, 97,* 346–353.

Weimer, W. B. (1979). *Notes on the methodology of scientific research.* Hillsdale, NJ: Erlbaum.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84,* 608–618.

Westen, D., Shedler, J., Durrett, C., Glass, S., & Martens, A. (2003). Personality diagnoses in adolescence: *DSM–IV* Axis II diagnoses and an empirically derived alternative. *American Journal of Psychiatry, 160,* 952–966.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Widiger, T. A., Costa, P. T., Jr., & McCrae, R. R. (2002). A proposal for Axis II: Diagnosing personality disorders using the five-factor model. In P. T. Costa, Jr. & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 431–456). Washington, DC: American Psychological Association.

Wood, J. M., Garb, H. N., Lilienfeld, S. O., & Nezworski, M. T. (2002). Clinical assessment. *Annual Review of Psychology, 53,* 519–543.

Wood, J. M., Nezworski, T. M., Lilienfeld, S. O., & Garb, H. N. (2003). *What's wrong with the Rorschach?* San Francisco: Wiley.

Zubin, J. (1954). Failures of the Rorschach technique. *Journal of Projective Techniques, 18,* 303–315.

Received August 5, 2003
Revision received October 19, 2004
Accepted November 3, 2004 ■