Construct Validity: Advances in Theory and Methodology

Milton E. Strauss[1] and Gregory T. Smith[2]

[1]Department of Psychology, University of New Mexico, Albuquerque, New Mexico 87131-0001; email: milton.strauss@gmail.com

[2]Department of Psychology, University of Kentucky, Lexington, Kentucky 40506-0044; email: gsmith@email.uky.edu

Annu. Rev. Clin. Psychol. 2009. 5:1–25

First published online as a Review in Advance on December 16, 2008

The *Annual Review of Clinical Psychology* is online at clinpsy.annualreviews.org

This article's doi: 10.1146/annurev.clinpsy.032408.153639

1548-5943/09/0427-0001$20.00

Key Words

philosophy of science, construct representation, multitrait-multimethod validation, construct homogeneity, construct validation programs

Abstract

Measures of psychological constructs are validated by testing whether they relate to measures of other constructs as specified by theory. Each test of relations between measures reflects on the validity of both the measures and the theory driving the test. Construct validation concerns the simultaneous process of measure and theory validation. In this article, we review the recent history of validation efforts in clinical psychological science that has led to this perspective, and we review the following recent advances in validation theory and methodology of importance for clinical researchers. These are: the emergence of nonjustificationist philosophy of science; an increasing appreciation for theory and the need for informative tests of construct validity; valid construct representation in experimental psychopathology; the need to avoid representing multidimensional constructs with a single score; and the emergence of effective new statistical tools for the evaluation of convergent and discriminant validity.

Contents

INTRODUCTION

In this review, we highlight the centrality of construct validation to theory testing in clinical psychology. In doing so, we first provide a brief history of modern validation efforts and describe the foundational role that construct validity theory has for modern, scientific clinical psychology. We then highlight four recent developments in construct validity theory and advances in statistical methodology that, we believe, should play an important role in shaping construct and theory validation efforts.

Construct: a psychological process or characteristic believed to account for individual or group differences in behavior

Construct validity: evaluation of the extent to which a measure assesses the construct it is deemed to measure

AN HISTORICAL OVERVIEW OF VALIDATION EFFORTS IN CLINICAL PSYCHOLOGY

At the modern beginning of scientific clinical psychology in the beginning of the twentieth century, researchers faced the challenge of developing valid measures without an existing knowledge base on which to rely. The absence of a foundation of knowledge was an enormous problem for test validation efforts. The goal of validating measures of psychological constructs necessarily requires criteria that are themselves valid. One cannot show that a predictor of some form of psychopathology is valid unless one can show that the predictor relates to an indicator of that form of psychopathology that is, itself, valid. One cannot show that a certain deficit in cognitive processing characterizes individuals with a certain disorder unless one has defined and validly measured the disorder. Inevitably, to validate scores on measures one needs a structure of existing knowledge to which one can relate those scores. To go further, to validate one's claim that scores on a measure play a certain role in a network of psychological processes one needs valid measures of the different components of the specified processes.

As researchers developed measures and confirmed or disconfirmed early, relatively crude predictive hypotheses, a knowledge base began to develop. The development of a knowledge base made possible the specification of procedures for measure validation. The specification of such procedures, in turn, facilitated further knowledge acquisition. And as knowledge continued to develop, the need for more theoretically sophisticated means of measure and theory validation emerged. We believe the recent history of validation efforts reflects this kind of reciprocal influence between existing knowledge and validation standards. We next briefly describe this process in greater detail.

Early Measure Development and Validity

An often-discussed early measure in the history of validation efforts is the Woodworth Personal

Data Sheet (WPDS), a measure developed in 1919 to help the U.S. Army screen out individuals who might be vulnerable to "war neurosis" or "shell shock." It was subsequently described as measuring emotional stability (Garrett & Schneck 1928, Morey 2002). Both during construction and use of the test, researchers showed clear concern with its validity. Unfortunately, their efforts to both develop and validate the test reflected the weak knowledge structure of clinical research at the time.

Woodworth constructed the 116-item test by relying on existing clinical psychological knowledge and by using empirical methods. Specifically, he drew his item content from case histories of individuals identified as neurotic. He then administered the items to a normal test group and deleted items scored in the presumably dysfunctional direction by 50% or more of that group (Garrett & Schneck 1928). Clearly, he sought to construct a valid measure of dysfunction. And although not all researchers who used the WPDS concerned themselves with its validity, some did. Flemming & Flemming (1929) chided researchers for neglecting to validate the test and then conducted their own empirical test of the measure.

Items on the WPDS are quite diverse. They include, "Have you ever lost your memory for a time?", "Can you sit still without fidgeting?", "Does it make you uneasy to have to cross a wide street or an open square?", and "Does some particular useless thought keep coming into your mind to bother you?" From the standpoint of today's knowledge base in clinical psychology, each of these four sample items seems to refer to a different construct. It is thus not surprising that the measure did not perform well. It did not differentiate college students from "avowed psychoneurotics" (Garrett & Schneck 1928), nor did it correlate with teacher ratings of students' emotional stability (Flemming & Flemming 1929).

One can see two core limitations underlying the effort to develop this test and validate it. First, in developing the WPDS item pool, Woodworth had to rely on a far too incomplete understanding of psychopathology and its contributors. Second, the validity of the criterion measures was not established independently and was based on either broad diagnostic classification or subjective teacher ratings; surely the validity of these criteria was limited.

Researchers at the time expressed concerns related to these limitations. For example, Garrett & Schneck (1928) noted the heterogeneous items and the mixture of complaints represented in the item pool and drew a conclusion (described in the Construct Homogeneity section below) that anticipated recent advances in validation theory:

> It is this [heterogeneity], among other [considerations], which is causing the present-day trend away from the concept of mental disease as an entity. Instead of saying that a patient has this or that disease, the modern psychiatrist prefers to say that the patient exhibits such and such symptoms (p. 465).

Based on this thinking, Garrett & Schneck (1928) investigated relations among individual items and specific diagnoses (rather than membership in the general category of "mentally disturbed"). In doing so, they recognized the need to avoid combining items of different content as well as the need to avoid combining individuals with different symptom pictures. Their use of an empirical item–person classification produced very different results from prior rational classifications (Laird 1925), thus (*a*) implicating the importance of empirical validation and (*b*) anticipating criterion-keying methods of test construction. In addition, they anticipated the current appreciation for construct homogeneity, with its emphasis on unidimensional traits and unidimensional symptoms as the preferred objects of theoretical study and measure validation (Edwards 2001, McGrath 2005, Smith et al. 2003, Smith & Combs 2008).

Construct homogeneity: the view that a single score should reflect variation on only a single construct

The Validation of Measures as Their Ability to Predict Criteria

During the early and middle parts of the twentieth century, test validity came to be understood

in terms of a test's ability to predict a practical criterion (Cureton 1950, Kane 2001). This focus on criterion prediction may have been a function of three forces: advances in substantive knowledge and precision of thought in the field, the obvious limitations in the tests constructed on purely rational grounds, and a philosophy-based suspicion of theories describing unobservable entities (Blumberg & Feigl 1931). Indeed, many validation theorists explicitly rejected the idea that scores on a test mean anything beyond their ability to predict an outcome. As Anastasi (1950) put it, "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).

At the time, this approach to measure validation proved quite useful: It led to the generation of new methods of test construction as well as to important substantive advances in knowledge. Concerning test construction, it led to the criterion-keying approach, in which one selects items entirely on the basis of whether the items predict the criterion. This method represented an important advance: To some degree, validity as successful criterion prediction was built into the test. The method worked well. Two of the most prominent measures of personality and psychopathology, the Minnesota Multiphasic Personality Inventory (MMPI; Butcher 1995) and the California Psychological Inventory (CPI; Megargee 2008), were developed using criterion keying. Each of those measures has generated a wealth of knowledge concerning personality, psychopathology, and adjustment: Thousands of studies attest to the measures' clinical value. For example, the MMPI-2 distinguishes between psychiatric inpatients and outpatients and facilitates treatment planning (Butcher 1990, Greene 2006, Nichols & Crowhurst 2006, Perry et al. 2006). It has also been applied usefully to normal populations (such as in personnel assessment; Butcher 2002, Derksen et al. 2003), to head-injured populations (Gass 2002), and in correctional facilities (Megargee 2006). The CPI validly predicts a wide range of criteria as well (Gough 1996).

As Kane (2001) noted, the criterion-related validity perspective also led to more sophisticated treatments of the relationship between test scores and criteria, as well as to the development of utility-based decision rules (see Cronbach & Gleser 1965). Perhaps it is also true that the focus on prediction of criteria as the defining feature of validity contributed to the finding that statistical combinations of test data are superior to clinical combinations, and that this is true across domains of inquiry (Grove et al. 2000, Swets et al. 2000).

As prediction improved and knowledge advanced using this criterion validity perspective, the ultimate limitations of the method became clear. One core limitation reflects a difficulty in prediction that was present from the beginning: tests of criterion-related validity are only as good as the criteria used in the prediction task. As Bechtoldt (1951) put it, reliance on criterion-related validity "involves the *acceptance* of a set of operations as an adequate definition of whatever is to be measured [or predicted]" (p. 1245). Typically, the validity of the criterion was presumed, not evaluated independently. In hindsight, there was good reason to question the validity of many criteria: They were often based on some form of judgment (crude diagnostic classification, teacher rating), and those judgments had to be made with an insufficiently developed knowledge base. Limitations in the validity of criteria impose limitations in one's capacity to validate a measure.

The second limitation is one that led to the development of construct validity theory and that could only have become apparent once the core knowledge base in clinical psychology had developed sufficiently: The criterion-related validity approach does not facilitate the development of basic theory. When tests are developed for the specific intent of predicting a circumscribed criterion, as is the case with criterion-keying test construction, and when they are only validated with respect to that predictive task, as is the case with criterion-related validity, the validation process is likely to contribute little to theory development. As a result, criterion-related validity findings tend not to

provide a strong foundation for deducing likely relationships among variables, and hence for the development of generative theory.

The Emergence of Construct Validity

In the early 1950s, there was an emerging concern with theory development that led to Meehl and Challman's introduction of the concept of construct validity in the 1954 Technical Recommendations (Am. Psychol. Assoc. 1954). Their work was part of the work of the American Psychological Association's Committee on Psychological Tests. In our view, the developing focus on theory was made possible, in part, by the substantive advances in clinical knowledge facilitated by the criterion-related validity approach. Perhaps ironically, the success of the criterion-related validity method led to its ultimate replacement with construct validity theory. The criterion approach led to significant advances in knowledge, which helped facilitate the development of integrative theories concerning cognition, personality, behavior, and psychopathology. But such theories could not be validated using the criterion approach; there was thus a need for advances in validation theory to make possible the emerging theoretical advances. This need was addressed by several construct validity authors in the middle of the twentieth century (Campbell & Fiske 1959, Cronbach & Meehl 1955, Loevinger 1957).

Indeed, theoretical progress in clinical psychology has substantially depended on four seminal papers, all published within a decade. The first (MacCorquodale & Meehl 1948) promoted the philosophical legitimacy of hypothetical constructs, concepts that have a "cognitive factual reference" (p. 107) that goes beyond the data used to support them. That is, hypothetical constructs are hypotheses about the existence of entities, processes, or events that are not directly observed. That seminal paper advanced the legitimacy of psychological theories that describe entities that underlie, but are not equivalent to, what is observed in the laboratory or other research setting.

The second paper (Cronbach & Meehl 1955) described the methods and rules of inference by which one develops evidence for the validity of measures of such hypothetical constructs. Construct validation tests are also tests of the validity of the theory that specifies a measure's presumed meaning. We use the word "developed" rather than "established" to emphasize that construct validation is an ongoing process, the process of theory testing. Central to Cronbach and Meehl's conceptualization of construct validity was the need to articulate specific theories describing relations among psychological processes, in order to then evaluate the performance of measures thought to represent one such process (see also Garner et al. 1956). Cronbach & Meehl (1955) emphasized deductive processes in construct validity. The third seminal paper (Loevinger 1957) identified the construct validation process as the general framework for the development and testing of psychological theories and the measures used to represent theoretical constructs. In Loevinger's view, construct validity subsumed both content validity and predictive/concurrent, or empirical, validity. In short, construct validity is validity (see also Landy 1986, Messick 1995).

The fourth paper (Campbell & Fiske 1959) considered issues in the validation of purported indicators of a construct. The title, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix (MTMM)," refers to two of the three core ideas in Fiske & Campbell's article that remain crucial in the process of validation of a measure as a construct indicator.

First, all measures are trait (construct)-method units. That is, variance in all psychological measures consists of the substantive construct variance, variance due to the method of measurement that is independent of the construct and, of course, errors of measurement. Second, two types of evidence are required to validate a test or other measurement in psychology. The first, convergent validity, is demonstrated by associations among "*independent* measurement procedures" designed to reflect the same or similar constructs

MTMM: multitrait-multimethod matrix. Method for evaluating the relative contribution of trait-related and method variance to the correlations among measures of multiple constructs

Convergent validity: the relationship among different measures of the same construct

Discriminant validity: demonstrations that a measure of a construct is unrelated to indicators of theoretically irrelevant constructs in the same domain

Method variance: the association among variables due to the similarity of operations in the measurement of these variables

(Campbell & Fiske 1959, p. 81; emphasis added). The second aspect of measurement validity, discriminant validity,[1] requires that a new measure of a construct be substantially less correlated with measures of conceptually unrelated constructs than with other indicators of that construct. Discriminant validity requires the contrast of relationships of measures of constructs in the same conceptual domain, e.g., personality or symptom dimension constructs. Although Campbell & Fiske (1959) gave even weight to convergent and discriminant validity, in later work, the initial primacy of convergent validity is acknowledged (Cook & Campbell 1979; see Ozer 1989). Third, because of the ever-present, often substantial method variance in all psychological measures, validation studies require the simultaneous consideration of two or more traits measured by at least two different methods. Campbell & Fiske (1959) referred to this approach as multitrait-multimethod matrix methodology; we return to this specific methodology at the end of this article.

Although these papers are over 50 years old, each remains an invaluable place to begin one's mastery of the concept of construct validity. From the first three of these foundational papers, we understand that each study using a measure is simultaneously a test of the validity of a measure and a test of the theory defining the construct. Each new test provides additional information supporting or undermining one's theory or validation claims; with each new test, the validity evidence develops further. Thus, validation is a process, not an outcome. Often, the construct validity of a measure is described as "demonstrated," which is incorrect (Cronbach & Meehl 1955). Although the process is ongoing, it is not necessarily infinite. For example, if a well-validated measure, such as the Wechsler Adult Intelligence Scale-III (Wechsler 1997) or the Positive and Negative Affect Scale (Watson et al. 1988b), does not behave as expected in a study, the measure would not be abandoned. One would likely retain one's confidence in the measure and consider other possible explanations for the outcome, such as deficient research design.

Since the time of these articles, it has also become clear that researchers should concern themselves with construct validity from the beginning of the test construction process. To develop a measure that validly represents a psychological entity, researchers should carefully define the construct and select items representing the definition (Clark & Watson 1995). This reasoning extends to the selection of parameters for manipulation in experimental psychopathology (see Knight & Silverstein 2001). As Bryant (2000) effectively put it for the assessment of a trait,

> Imagine, for example, that you created an instrument to measure the extent to which an individual is a "nerd." To demonstrate construct validity, you would need a clear initial definition of what a nerd is to show that the instrument in fact measures "nerdiness." Furthermore, without a precise definition of nerd, you would have no way of distinguishing your measure of the nerdiness construct from measures of shyness, introversion or nonconformity (p. 112).

There have been four recent developments in perspectives on construct validity theory of importance for clinical psychological measurement. First, the philosophical understanding of scientific inquiry has evolved in ways that underscore both the complexity and the indeterminate nature of the validation process (Bartley 1987, Weimer 1979). Second, it has become apparent that the relative absence of strong, precise theories in clinical psychology sometimes leads to weak, noninformative validation tests (Cronbach 1988, Kane 2001). Appreciation of this has led theorists to reemphasize the centrality of theory testing in construct validation (Borsboom et al. 2004, Kane 2001). Third, researchers have accentuated the need to consider, as an aspect of construct validity, evaluation of theories describing the psychological

[1]Discriminant validity is sometimes erroneously referred to as divergent validity.

processes that lead to responses in psychological experiments such as are used in experimental psychopathology research. Tests of such theories are evaluations of construct representation [Whitely (now Embretson) 1983, Embretson 1998; see Knight & Silverstein 2001]. Fourth, researchers have stressed the importance of specifying and measuring homogeneous constructs, so the meaning of validation tests is unambiguous (Edwards 2001; Hough & Schneider 1995; McGrath 2005; Schneider et al. 1996; Smith et al. 2003; Smith & McCarthy 1995; G.T. Smith, D.M. McCarthy, T.B. Zapolski, manuscript submitted). We consider each of these in turn. But first, what is the current view of construct validity in assessment?

Current Views on Construct Validity in Psychological Measurement

Construct validity is now generally viewed as a unifying form of validity for psychological measurements, subsuming both content and criterion validity, which traditionally had been treated as distinct forms of validity (Landy 1986). Messick (1989, as discussed in Messick 1995) has argued that even this notion of validity is also too narrow. In his view, "[v]alidity is an overall evaluative judgment of the degree to which [multiple forms of] evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores. . ." (Messick 1995, p. 741).

That is, construct validity is comprehensive, encompassing all sources of evidence supporting specific interpretations of a score from a measure as well as actions based on such interpretations. Messick, writing mainly with reference to educational assessment, identified six contributors to construct validity (Messick 1995, see Figure 1, p. 748): (1) content relevance and technical quality; (2) theoretical understanding of scores and associated empirical evidence, including process analyses; (3) structural data; (4) generalizability; (5) external correlates; and (6) consequences of score interpretation. We focus here on aspects 2, 3, and 5, considering points 1 and 4 to be relatively well established and not controversial, and the practical consequence of test use (point 6) to be beyond the scope of this chapter (but see Youngstrom 2008).

ADVANCES IN PHILOSOPHY OF SCIENCE

In the first half of the twentieth century, many philosophers of science held the view that theories could be fully justified or fully disproved based on empirical evidence. The classic idea of the critical experiment that could falsify a theory is part of this perspective, which has been called justificationism (Bartley 1962, Duhem 1914, 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. From this perspective, 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, advances in the philosophy of science have led to a convergence on a different perspective, referred to as nonjustificationism (Bartley 1987; Campbell 1987, 1990; Feyerabend 1970; Kuhn 1970; Lakatos 1968; Weimer 1979). The nonjustificationist perspective is that no theory is ever fully proved or disproved. Instead, in the ongoing process of theory development and evaluation, at a given time certain theories are viewed as closer approximations to truth than are other theories. From this perspective (which dominates current philosophy of science, despite disagreement both within and outside this framework; Hacking 1999, Kusch 2002, Latour 1999), science is understood to be characterized by a lack of certainty.

The reason for the uncertainty is as follows. When one tests any theory, such as "individual differences in personality cause individuals to react differently to the same stimulus" (a theory of considerable importance for understanding the process of risk for psychopathology; Caspi 1993), one is presupposing the

Construct representation: the analysis of psychological processes accounting for responses on a task

Nonjustificationist: the philosophy of science that proposes that no theory is ever fully proven or disproven; rather, theories are selected on the basis of which one of several the bulk of evidence favors

validity of multiple theories in order to conduct the test (Lakatos 1999; Meehl 1978, 1990). In this example, one must accept that (*a*) there are reliable individual differences in personality that are not fully a function of context; (*b*) one has measured the appropriate domains of individual differences in personality; (*c*) one's measure of personality is valid, in that variation on dimensions of personality underlie variation in responses to the measure; (*d*) one's measure of personality does not represent other, nonpersonality processes to any substantial degree; (*e*) one's measure of each specific dimension of personality is coherent and unidimensional, i.e., does not represent variation on multiple dimensions simultaneously; (*f*) one can validly expose different individuals to precisely the same stimulus; (*g*) one can validly measure reactions to that stimulus; and so on.

It is easy to see that a failed test of the initial, core hypothesis could actually be due not just to a failure of the theory, but instead to failures in any number of "auxiliary" theories invoked to test the hypothesis. Researchers typically consider a number of different possibilities when faced with a nonsupportive finding. Often, when one faces a negative finding for a theory one believes has strong support otherwise, one questions any number of auxiliary issues: measurement, sample, context, etc. Doing so is quite appropriate (Cronbach & Meehl 1955).

Science is characterized by ongoing debates between proponents and opponents of a theoretical perspective. Through the ongoing process of theoretical criticism and new empirical findings, the debate comes to favor one side over the other. In considering this process, Weimer (1979) concluded that what characterizes science is "comprehensively critical rationalism" (p. 40), by which he meant that every aspect of the research enterprise must be open to criticism and potential revision. Researchers must make judgments as to whether one should question a core theory, an auxiliary theory, or both; they must then investigate the validity of those judgments empirically.

Thus, validation efforts can be understood as arguments concerning the overall evaluation of the claimed interpretation of test scores (Messick 1995) or of claims concerning the underlying theory (Kane 2001). The validation enterprise can thus be understood to include a coherent analysis of the evidence for and against theory claims. Researchers can design theory validation tests based on their analysis of the sum total of evidence relevant to the target theory.

Interestingly, this perspective, particularly as argued by psychological scientists, has begun to influence inquiry in historically nonempirical fields as well. For example, legal scholars, drawing on construct validation theory, have begun to argue that empirical investigation of legal arguments is a necessary part of the validation of those theories (Goldman 2007). Their contention is that sound arguments for the validity of legal theories require both theoretical coherence and supportive empirical evidence.

There is no obvious answer to the question of how one decides which theoretical arguments, embodied by programs of research, are convincing and which are not. Lakatos (1999) referred to progressing versus degenerating research programs. Progressing research programs predict facts that are subsequently confirmed by research; degenerating research programs may offer explanations for existing findings, but they do not make future predictions successfully, and they often require post hoc theoretical shifts to incorporate unanticipated findings (Lakatos 1999). Clearly, this perspective requires judgment on the part of researchers.

It is important to appreciate that the concept of the nonjustificationist nature of scientific inquiry did not spring from studies of psychology as a science. Most authors espousing these views have focused primarily on hard sciences, particularly physics. It is a reality of scientific inquiry that findings are always open to challenge and critical evaluation. Indeed, what separates science from other forms of inquiry is that it embraces critical evaluation, both by

theory and by empirical investigation (Weimer 1979). A second point is equally important to appreciate: The reality that no theories are ever fully proved or disproved is no excuse to proceed without theory or without clearly articulated theoretical predictions.

STRONG, WEAK, AND INFORMATIVE PROGRAMS OF CONSTRUCT VALIDATION

As discussed by Kane (2001), there have been drawbacks in the use of construct validity theory to organize measure and theory validation. The core idea that one can define constructs by their place in a lawful network of relationships (the network is deduced from the theory) assumes a theoretical precision that tends not to be present in the social sciences. Typically, clinical psychology researchers are faced with the task of validating their measures and theories despite the absence of a set of precisely definable, expected lawful relations among construct measures. Under this circumstance, the meaning of construct validity, and what counts as construct validation, is ambiguous.

Cronbach (1988) addressed this issue by contrasting strong and weak programs of construct validity. Strong programs depend on precise theory and are perhaps accurately understood to represent an ideal. Weak programs, on the other hand, stem from weak, or less fully articulated, theories and construct definitions. With weak validation programs, there is less guidance as to what counts as validity evidence (Kane 2001). One result can be approaches in which almost any correlation can be described as validation evidence (Cronbach 1988). In the absence of a commitment to precise construct definitions and specific theories, validation research can have an ad hoc, opportunistic quality (Kane 2001), the results of which tend not to be very informative.

Informative, Rather Than Strong or Weak, Theory Tests

In our view, clinical researchers are not wedged between a yet unattainable ideal of strong theory and ill-conceived, weak theory testing. Rather, there is an iterative process in which tests of partially developed theories provide information that leads to theory refinement and elaboration, which in turn provides a sounder basis for subsequent construct and theory validation research. Cronbach & Meehl (1955) referred to this bootstrapping process and to the inductive quality of construct definition and theory articulation; advances in testing partially formed theories lead to the development of more elaborate, complete theories. This process has proven effective; striking advances in clinical research have provided clear benefits to the consumers of clinical services.

DBT: dialectical behavior therapy

One example of this process has been the development of an effective psychological treatment for many of the behaviors characteristic of a previously untreatable disorder: borderline personality disorder. Dialectical behavior therapy (DBT) provides improved treatment of parasuicidal behavior and excessive anger (Linehan 1993, Linehan et al. 1993). The emergence of this treatment depended on incremental advances in numerous domains of clinical inquiry. First, advances in temperament theory and personality theory led to awareness of the stability of human temperament and personality, even across decades (Caspi & Roberts 2001, Roberts & DelVecchio 2000). That finding carried the obvious implication that treatment aimed at altering personality may not prove effective. The second advance was the recognition of disorders of personality, i.e., chronic dysfunction in characteristic modes of thinking, perceiving, and behaving, as distinct from other sources of dysfunction (Millon et al. 1996). That recognition facilitated the emergence of treatments targeted toward one's ongoing, typical mode of reacting and behaving. The third advance was the finding that behavioral interventions were effective for disorders of mood: When depressed individuals began participating in numerous, previously rewarding activities, their mood altered (Dimidjian et al. 2006).

DBT can be understood to represent the fruitful integration of each of these three theoretical advances. DBT was designed to treat

individuals with borderline personality disorder. One central aspect of DBT is that therapists do not attempt to change borderline clients' characteristic, intense affective response style: Attempts to do so are unlikely to be successful, given the stability of personality. Instead, therapists seek to provide behavioral skills for clients to employ to manage their intense affective reactivity. The therapeutic focus has become managing one's mood effectively, and it has proven effective (Linehan 1993).

To facilitate the process of theory development, researchers should consider whether their theoretical statements and tests are informative, given the current state of knowledge (Smith 2005). Is a theory consistent with what else is known in the field (MacCorquodale & Meehl 1948)? Can it be integrated with existing knowledge? To what degree does a hypothesis test shed light on the likely validity of a theory, or the likely validity of a measure? Does a hypothesis involve a direct comparison between two alternative theoretical explanations? Does a hypothesis involve a claim that, if supported, undermines criticism of a theory? Does a hypothesis involve direct criticism of a theory, or a direct challenge to the validity of a measure? Theory tests of this kind will tend to advance knowledge because they facilitate the central component of the scientific process: critical evaluation and cumulative knowledge.

Recent Arguments for a Reconceptualization of the Role of Theory in Clinical Research

In recent years, validity theorists have argued for an increased emphasis on theory in several aspects of psychological inquiry (Barrett 2005; Borsboom 2006; Borsboom et al. 2003, 2004; Maraun & Peters 2005; Michell 2000, 2001). We next review three basic arguments offered in this recent writing; we believe two of these apply, straightforwardly, to clinical science, and the third does not.

The first argument, which we consider both relevant to clinical research and uncontroversial, concerns latent variable theory. Latent variable theory reflects the idea that variation in responses to test items indicates variation in levels of an underlying trait. As Borsboom et al. (2003) most recently noted, latent variable theory involves a specific view of causality: Variation in a construct causes variation in test scores. When clinical psychology researchers describe a scale as a valid measure of a construct, such as anxiety, they are saying that variation in anxiety among individuals causes variation in those individuals' test responses. From this point of view, each item on a test is an indicator of the construct of interest. Borsboom et al. (2003) develop the implications of this theory for psychological assessment.

The second argument concerns the basic distinction between theory and empirical data: Theories exist independently of data (Borsboom 2006). It is certainly appropriate for researchers to develop, adopt, and promote explicit theories of psychological processes. Of course, ideally, researchers avoid inferring that findings provide stronger support for theories than they do, but that appropriate caution should not dissuade researchers from taking clear theoretical stands. More explicit statements of theory would (*a*) clarify the degree to which a given empirical test truly pertains to the theory and (*b*) drive the development of more direct tests of theoretical mechanisms (Borsboom 2006, Borsboom et al. 2004).

The third recent argument is one that, we believe, does not accurately pertain to the development of clinical science. Several authors emphasize the need for more explicit, well-developed theories in general (Barrett 2005; Borsboom 2006; Maraun & Peters 2005; Michell 2000, 2001). At least one of these writers (Borsboom 2006) emphasizes the need to begin with precise, fully developed theories; in Borsboom's view, to do otherwise is to provide a disservice to the field. For example, although psychological theories often refer to causal processes, they are neither detailed nor mathematically formal. From the point of view of these authors, this is regrettable.

This point of view has not gone without criticism. Both Clark (2006) and Kane (2006)

note that the incomplete knowledge base in psychology requires that any theory be an approximation, to be modified as new knowledge becomes available. Formal mathematical theories of psychological phenomena, especially in clinical psychology, are quite premature. And regardless of how detailed and precise the explication of a theory is, each component of it would necessarily undergo critical evaluation and revision as part of the normal progress of science (Weimer 1979). It seems to us that this process is inevitable and is a normal part of scientific inquiry.

Construct Representation and Nomothetic Span

Whitely (1983, Embretson 1998) introduced an important distinction in construct validity theory between nomothetic span and construct representation. Nomothetic span refers to the pattern of significant relations among measures of the same or different constructs (i.e., convergent and discriminant validity). Nomothetic span is in the domain of individual differences (correlation). It is particularly relevant to research concerning expected relationships among trait measures or measures of intellectual skills, neuropsychological variables, or measures of personality constructs. For example, IQ has excellent nomothetic span because individual differences in various measures of that construct all show similar meaningful patterns of relationship with other variables as expected (Whitely 1983). A method for evaluating nomothetic span is the confirmatory factor analysis of a matrix of correlations among measures for which there are specifications of which relationships should be present and which should be absent.

Construct representation (Embretson 1998, Whitely 1983), on the other hand, refers to the validation of the theory of the response processes that result in a score (such as accuracy or reaction time) in the performance of cognitive tasks. That is, construct representation refers to the psychological processes that lead to a given response on a trial or to the pattern of responses across conditions in an experiment. For many authors, and particularly for cognitive psychologists, construct representation indicates the validity of the dependent variable as an index of a construct (Borsboom et al. 2004, Embretson 1998). That is to say, the goal of construct representation is to test a theory of the cognitive processes giving rise to a response.

Nomothetic span: the meaning of a construct as established through its network of relationships with other constructs

An example may make the notion of construct representation clearer. Carpenter et al. (1990) proposed a theory of matrix reasoning problem solving to account for performance on tests such as Ravens Progressive Matrix test, a widely used measure of intelligence. Their model posited that working memory was a critical determinant of performance and that working memory load is influenced by two parameters: (*a*) the number of relationships among elements in a matrix and (*b*) the level of complexity of the relationships among the elements. Note that these are quantitative variables. So by developing matrix items that systematically varied on these two dimensions, these investigators were able to evaluate the extent to which each parametric variation separately and conjointly determined performance.

The model, in other words, identified the underlying psychological processes that were validated, through accounting for performance on the task as the proposed processes were parametrically manipulated. The validity of the model provides evidence of the construct representation component of the test. The Ravens thus has both evidence of construct representation (model predictions are confirmed) and nomothetic span in that individual differences in performance on the standardized version of the test correlate meaningfully with other variables. Nomothetic span and construct representation aspects of construct validity can complement each other. As an example, the construct representation analysis of Carpenter et al. (1990) is supported by correlational analyses showing that working memory tests but not tests of short-term memory are related to measures of fluid intelligence (Engle et al. 1999).

On the other hand, measures may have developed evidence of construct validity of one

sort but not the other. Most IQ measures have excellent nomothetic span but limited construct representation: Scores predict many things, but the specific psychological processes underlying responses (and those underlying processes common across measures) are generally unknown. The converse may also be true. As Whitely (1983) describes, Posner's (1978) verbal encoding task has excellent construct representation: The psychological mechanisms underlying performance are well established. However, the task has poor nomothetic span because individual differences on that task do not correlate well with scores on other measures of verbal encoding (Whitely 1983).

Construct Representation Research in Clinical Psychology

Construct representation has been understudied in clinical psychology research, particularly in clinical neuropsychology and experimental psychopathology. Theories of schizophrenia, depression, and other disorders emphasize disruptions in cognitive processes, and the nomothetic span of a number of tests within neuropsychology, cognitive psychology, and clinical cognitive neuroscience paradigms are well established. But the construct representation of such tests is often less well developed: Many are psychologically complex, many are adaptations of paradigms developed for studying normal cognition, and at least in the case of schizophrenia research, many are poorly understood in terms of the underlying processes accounting for task deficits (Strauss 2001, Strauss & Summerfelt 1994). How construct representation may be relevant to research with personality or symptom self-reports or interviews is unclear and is a topic for further conceptual analysis and research.

Although construct representation and nomothetic span are distinct, one can influence the other. Performance on cognitive and neuropsychological tasks involves the operation of multiple cognitive processes, each of which may be reliably measured by the task. However, some of the reliable variance may well be construct irrelevant (Messick 1995, Silverstein 2008). In such instances, group differences on a task as well as associations between task performance and conceptually relevant other variables (i.e., apparent nomothetic span) may be due to such reliable but construct-irrelevant variance (Messick 1995, Silverstein 2008). Theoretical progress in clinical cognitive neuroscience and experimental psychopathology depend on the conjoint analysis of nomothetic span and construct representation in the evaluation of the construct validity of measures.

Conjoint analysis of nomothetic span and construct representation is also important for theory development in the study of personality traits and symptoms, especially as the field becomes more focused on neurobiological processes in personality and psychopathology. For example, there are at least 27 studies of the relation of impulsivity to the Iowa Gambling Task, a proposed measure of neurobiologically based deficits in decision making (PsychInfo search, July 1, 2008, with terms "Iowa Gambling Task" and "impulsivity"). However, none of these studies has evaluated the construct representation of the task, which is necessary to develop links between neurobiology, psychological processes, and individual differences in impulsivity. An excellent example of the conjoint evaluation of construct representation and nomothetic span is the work of Richard DePue, who has proposed a detailed theory of the biology of extraversion and its link with psychopathology (e.g., Depue & Collins 1999).

The incorporation of converging operations (Garner et al. 1956) into research designs can facilitate the analysis of construct representation and identify the extent to which correlations between performance and other variables reflect construct-relevant associations. For clinical research, the ability of different tasks or individual-difference measures to differentially predict markers of observable, clinically important behaviors speaks to the presence of substantial construct-relevant variance (Hammond et al. 1986).

Establishing the construct representation of a measure requires an explicit theoretical

analysis of test performance and empirical tests of the theory (Whitely 1983). An example of such a research program is the experimental analysis of the basis of schizophrenia patients' error patterns on the A-X CPT, a form of vigilance task widely used in schizophrenia research (see Cornblatt & Keilp 1994, Nuechterlein 1991). In the A-X CPT, subjects must respond to the brief occurrence of an X in a rapidly changing sequence of letters, but only if the X is preceded by an A (Cohen et al. 1999). Experiments evaluating a theory of construct representation in this task suggested deficits in context representation as the most fruitful interpretation of task performance. A number of experiments using converging operations along with manipulations of theoretically proposed constituent processes have converged on this conclusion (see Barch 2005, Cohen et al. 1999). There is also substantial evidence of nomothetic span validity for the A-X CPT, including the specificity of the deficit to schizophrenia among psychotic disorders, as well as association with specific symptoms, intellectual function, and genetic liability to schizophrenia spectrum disorders (Barch et al. 2003, MacDonald et al. 2005). Other research programs suggest that this deficit may be an instance of a more general deficit in contextual coordination at both the behavioral and neural levels (Phillips & Silverstein 2003, Uhlhaas & Silverstein 2005).

CONSTRUCT HOMOGENEITY

Over the past 10 to 15 years, psychometric theory has evolved in a fundamental way that is crucial for psychopathology researchers to appreciate. In the past, psychometrics writers argued for the importance of including items on scales that tap a broad range of content. Researchers were taught to avoid including items that were highly redundant with each other, because then the breadth of the scale would be diminished and the resulting high reliability would be associated with an attenuation of validity (Loevinger 1954). To take the logic further, researchers were sometimes encouraged to choose items that were largely uncorrelated with each other, so that each new item could add the greatest possible incremental predictive validity over the other items (Meehl 1992).

NEO-PI-R: NEO Personality Inventory-Revised

In recent years, a number of psychometricians have identified a core difficulty with this approach. If items are only moderately intercorrelated, it is likely that they do not represent the same underlying construct. As a result, the meaning of a score on such a test is unclear. Edwards (2001) noted that researchers have long appreciated the need to avoid heterogeneous items: If such an item predicts a criterion, one will not know which aspect of the item accounts for the covariance. The same reasoning extends to tests: If one uses a single score from a test with multiple dimensions, one cannot know which dimensions account for the test's covariance with measures of other constructs.

There are two sources of uncertainty built into any validation test that uses a single score to reflect multiple dimensions. The first is that one cannot know the nature of the different dimensions' contributions to that score and hence to correlations of the measure with measures of other constructs. The second source of uncertainty is perhaps more severe than the first. The same composite score is likely to reflect different combinations of constructs for different members of the sample.

McGrath (2005) clarified this point by drawing a useful distinction between psychological constructs that represent variability on a single dimension, on the one hand, and concepts designed to refer to covariations among unidimensional constructs, on the other hand. Consider the NEO Personality Inventory-Revised (NEO-PI-R) measure of the five-factor model of personality (Costa & McCrae 1992). One of the five factors is neuroticism, which is understood to be composed of six elemental constructs. Two of those are angry hostility and self-consciousness. Measures of those two traits covary reliably; they consistently fall on a neuroticism factor in exploratory factor analyses conducted in different samples and across cultures (McCrae et al. 1996). However, they are not the same construct. Their correlation was

0.37 in the standardization sample; they share only 14% of their variance. When concerned with theoretical issues it is appropriate to disattenuate correlations for unreliability. In this instance, the common variance between angry hostility and self-consciousness, corrected for unreliability, is estimated to be 19%.

Clearly, one person could be high in angry hostility and low in self-consciousness, and another could be low in angry hostility and high in self-consciousness. Those two different patterns could produce exactly the same score on neuroticism as measured by the NEO-PI-R, even though the two traits may have importantly different correlates. For example, the consensus view of psychopathy, based on both expert ratings and measurement, involves being unusually high in angry hostility and unusually low in self-consciousness (Lynam & Widiger 2007). Thus, it makes sense to develop theories relating angry hostility, or self-consciousness, to other constructs, and tests of such theories would be coherent. However, a theory relating overall neuroticism to other constructs must be imprecise and unclear because of the relative independence of the facets of the construct. If neuroticism correlates with another measure, one does not know which traits account for the covariation, or even whether the same traits account for the covariation for each member of the sample.

The use of a neuroticism score, obtained as a summation of scores on several, separable traits, is problematic because it introduces theoretical imprecision. That observation is separate from the theoretical claim that there is a unidimensional construct, whether referred to as negative affectivity or emotional instability, which relates to variability on each lower-level construct within the broad neuroticism domain. There is, of course, considerable empirical support for that claim (Costa & McCrae 1992, Watson et al. 1988a), as well as support for the view that each lower-level construct shares variance with general negative affectivity and also has variance specific to the lower-level construct (Krueger et al. 2001). We note that since the specific variance for each lower-level construct can be substantial, summing scores on the lower-level constructs to obtain a single overall score introduces theoretical and empirical imprecision as we described above.

Hough & Schneider (1995), McGrath (2005), Paunonen & Ashton (2001), Schneider et al. (1996), and Smith et al. (2003), among others, have noted that the use of scores of broad measures often obscures predictive relationships. Paunonen (1998) and Paunonen & Ashton (2001) have shown that prediction of theoretically relevant criteria is improved when one uses facets of the Big Five personality scales, rather than the composite, big five dimensions themselves. Using the NEO-PI-R operationalization of the five-factor model of personality, Costa & McCrae (1995) compared different facets of conscientiousness in their prediction of aspects of occupational performance. Dutifulness was related to service orientation (0.35) and employee reliability (0.38), but achievement striving was not (–0.01 and 0.02, respectively). In contrast, achievement striving was related to sales potential (0.22), but dutifulness was not (0.06). By definition, correlations of broad conscientiousness (which on the NEO-PI-R sums these two facets with four other facets) will produce correlations in between the high and low values because the sum effectively averages the different effects of the different facets. Use of the broad score would obscure the different roles of the different facets of conscientiousness. Should one wish to represent the full domain of a higher-order dimension, such as conscientiousness or neuroticism, one can include each lower-level facet as part of a multivariate analysis (such as multiple regression); doing so preserves the theoretical precision inherent in precise constructs while representing the full variance of the higher-order domain (Goldberg 1993, Nunnally & Bernstein 1994).

Recently, this perspective has been extended to the study of disorders. For example, McGrath (2005), noting that individuals can obtain the same depression scores with very different symptom patterns, describes depression as a useful social construction but not a

coherent psychological entity that can be used in validation studies. Indeed, using factor analysis, Jang et al. (2004) identified 14 subfactors in a set of depression measures. Examples included "feeling blue and lonely," "insomnia," "positive affect," "loss of appetite," and "psychomotor retardation." They found that the intercorrelations among the factors ranged from 0.00 to 0.34; further, the factors were differentially heritable, with heritability coefficients ranging from 0.00 to 0.35. Evidence of multidimensionality is accruing for many disorders, including posttraumatic stress disorder (King et al. 1998, Simms et al. 2002), psychopathy (Brinkley et al. 2004), schizotypal personality disorder (Fossati et al. 2005), and many others (Smith & Combs 2008).

For scientific clinical psychology to advance, researchers should study cohesive, unidimensional constructs. To use multifaceted, complex constructs as predictors or criteria in validity or theory studies is difficult to defend. Researchers are encouraged to generate theories that identify putatively homogenous, coherent constructs. It may often be useful to compare the theory that a putative attribute is homogeneous to the theory that it is a combination of separate attributes. The success of such efforts in the recent past bodes well for continued progress in the field as researchers study unidimensional constructs with meaningful test scores (Jang et al. 2004, Smith et al. 2007, Whiteside & Lynam 2001).

This discussion of construct homogeneity raises two important issues. The first is, when is a construct measure elemental enough? There is a risk of continually parsing constructs until one is left with a content domain specific to a single item, thus losing full coverage of a target construct and attenuating predictive power. We believe the guiding consideration should be theoretical clarity. When there is good theoretical or empirical reason to believe that an item set actually consists of two separately definable constructs with different psychological meaning, and when those two constructs could reasonably have meaningfully different external correlates, measuring the two separately is likely to improve both understanding and empirical prediction. When there is no sound theoretical basis to separate items into multiple constructs, one should perhaps avoid doing so.

The second issue is whether a focus on construct homogeneity leads to a clear and unacceptable loss of parsimony. This possibility merits careful consideration. With respect to etiological models, the use of several homogeneous constructs rather than their aggregate can complicate theory testing, but that difficulty must be weighed against the improved precision of theory tests.

It is at least possible that an emphasis on construct homogeneity often does not compromise parsimony. For example, it appears to be the case that four broad personality dimensions and their underlying facets effectively describe the many different forms of dysfunction currently represented by the full set of personality disorders (Widiger & Simonsen 2005, Widiger et al. 2005). Perhaps it is instead the case that parsimony has been compromised by the current *Diagnostic and Statistical Manual* system that names multiple putative syndromes that often appear to reflect slightly different combinations of personality dimensions.

It may be that parsimony would be better served by describing personality dysfunction in terms of a set of core, homogeneous personality traits rather than in terms of combinations of disparate, moderately related symptoms (Widiger & Trull 2007). This logic has been extended beyond the personality disorders domain: Serretti & Olgiati (2004) described basic dimensions of psychosis that apply across current diagnostic distinctions, suggesting parsimony in the dimensional description of psychosis.

EMPIRICAL EVALUATION OF CONSTRUCT VALIDITY

Campbells & Fiske's (1959) multitrait-multimethod matrix methodology presented a logic for evaluating construct validity through simultaneous evaluation of convergent and discriminant validity, and the contribution of

method variance to observed relationships.[2] Wothke (1995) nicely summarized the central idea of MTMM matrix methodology:

> The crossed-measurement design in the MTMM matrix derives from a simple rationale: Traits are universal, manifest over a variety of situations, and detectable with a variety of methods. Most importantly, the magnitude of a trait should not change just because different assessment methods are used (p. 125).

Traits are latent variables, inferred constructs. The term "trait," as used here, is not limited to enduring characteristics; it also applies to more transitory phenomena such as moods and emotions, as well as to all other individual differences constructs, e.g., attitudes and psychophysical measurements. Methods for Campbell and Fiske are the procedures through which responses are obtained, the operationalization of the assessment procedures that produce the responses, the quantitative summary of which is the measure itself (Wothke 1995).

As Campbell & Fiske (1959) emphasized, measurement methods (method variance) are sources of irrelevant, though reliable, variance. When the same method is used across measures, the presence of reliable method variance can lead to an overestimation of the magnitude of relations among constructs. This can lead to overestimating convergent validity and underestimating discriminant validity. This is why multiple assessment methods are critical in the development of construct validity. Their distinction of validity (the correlation between dissimilar measures of a characteristic) from reliability (the correlation between similar measures of a characteristic) hinged on the differences between construct assessment methods.

Campbell & Fiske's (1959) observation remains important today: Much clinical psychology research relies on the same method for both predictor and criterion measurement, typically self-report questionnaire or interview. Their call for attention to method variance is as relevant today as it was 50 years ago; examination of constructs with different methods is a crucial part of the construct validation process. Of course, the degree to which two methods are independent is not always clear. For example, how different are the methods of interview and questionnaire? Both rely on self-report, so are they independent sources of information? Perhaps not, but they do differ operationally. For example, questionnaire responses are often anonymous, whereas interview responses require disclosure to another. Questionnaire responses are based on the perceptions of the respondent, whereas interview ratings are based, in part, on the perceptions of the interviewer. A conceptually based definition of "method variance" has not been easy to achieve, as Sechrest et al.'s (2000) analysis of this issue demonstrates. Certainly, method differences lie on a continuum where, for example, self-report and interview are closer to each other than are self-report and informant report or behavioral observation.

The guidance provided for evaluating construct validity in 1959 was qualitative; it involved the rule-based examination of patterns of correlations against the expectations of convergent and discriminant validity (Campbell & Fiske 1959). Developments in psychometric theory, multivariate statistics, and analysis of latent traits in the decades since the Campbell & Fiske (1959) paper have made available a number of quantitative methods for modeling convergent and discriminant validity across different assessment methods.

Bryant (2000) provides a particularly accessible description of using analysis of variance (and a nonparametric variant) and confirmatory factor analysis (CFA) in the analysis of MTMM matrices. A major advantage of CFA

[2]Campbell and Fiske were concerned with correlation designs relevant to research on traits and other individual differences constructs. Similar issues apply to experimental psychology paradigms, as Garner et al. (1956) described in their discussion of convergent operations. This article is particularly relevant to experimental psychopathology, where multiple paradigms for the study of cognitive impairments in patients are particularly important for the identification of disordered cognitive mechanisms, structures, or processes (e.g., see Knight & Silverstein 2001).

in construct validity research is the possibility of directly comparing alternative models of relationships among constructs, a critical component of theory testing (see Whitely 1983). Covariance component analysis of the MTMM matrix has also been developed (Wothke 1995). Both covariance component analysis and CFA are variants of structural equation models (SEMs). With these advances, eyeball examinations of MTMM matrices are no longer sufficient for the evaluation of the trait validity of a measure in modern assessment research.

Perhaps the first CFA approach was one that followed very straightforwardly from Campbell & Fiske (1959): It involved specifying a CFA model in which 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). So if traits A, B, and C are each measured with methods X, Y, and Z, there are six latent variables: three for the traits and three for the methods. Thus, if indicator *i* reflects method X for evaluating trait A, that part of the variance of *i* that is shared with other indicators of trait A is assigned to the trait A factor, that part of the variance of *i* that is shared with indicators of other constructs measured by method X is assigned to the method X 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 test explicitly the role of a certain trait or a certain type of method variance on responses to a criterion measure. This approach can be expanded to include interactions between traits and methods (Campbell & O'Connell 1967, 1982), and therefore test multiplicative models (Browne 1984, Cudeck 1988).

Although the potential advantages of this approach are obvious, it has generally not proven feasible. As noted by Kenny & Kashy (1992), this approach often results in modeling more factors than there is information to identify them; the result often is a statistical failure to converge on a factor solution. That reality has led some researchers to turn away from multivariate statistical methods to evaluate MTMM results. In recent years, however, two alternative CFA modeling approaches have been developed that appear to work well.

SEMs: structural equation models

The first is referred to as the "correlated uniquenesses" approach (Marsh & Grayson 1995). In this approach, one does not model method factors as in the approach previously described. Instead, one identifies the presence of method variance by allowing the residual variances of trait indicators that share the same method to correlate, after accounting for trait variation and covariation. To the degree there are substantial correlations between these residual terms, method variance is considered present and is accounted for statistically (although other forms of reliable specificity may be represented in those correlations as well). As a result, the latent variables reflecting trait variation do not include that method variance: One can test the relation between method-free trait scores and other variables of interest. And, since this approach models only trait factors, it avoids the overfactoring problem of the earlier approach. There is, however, an important limitation to the correlated uniquenesses approach. Without a representation of method variance as a factor, one cannot examine the association of method variance with other constructs, which may be important to do (Cronbach 1995).

The second alternative approach provides a way to model some method variance while avoiding the overfactoring problem (Eid et al. 2003). One constructs latent variables to represent all trait factors and all but one method factor. Since there are fewer factors than in the original approach, the resulting solution is mathematically identified: One has not overfactored. The idea is that one method is chosen as the baseline method and is not represented by a latent variable. One evaluates other methods for how they influence results compared to the baseline method. Suppose, for example, that one had interview and collateral report data for a series of traits. One might specify the interview method as the baseline method, so an interview method factor is

not modeled as separate from trait variance, and trait scores are really trait-as-measured-by-interview scores. One then models a method factor for collateral report. If the collateral report method leads to higher estimates of trait presence than does the interview, one would find that the collateral report method factor correlated positively with the trait-as-measured-by-interview. That would imply that collaterals report higher levels of the trait than do individuals during interviews.

Interestingly, one can assess whether this process works differently for different traits. Perhaps collaterals perceive higher levels of some traits than are reported by interview (unflattering traits?) and lower levels of other traits as reported by interview (flattering traits?). This possibility can be examined empirically using this method. In this way, the Eid et al. (2003) approach makes it possible to identify the contribution of method to measure scores. The limitation of this method, of course, is that the choice of baseline method influences the results and may be arbitrary (Eid et al. 2003).

Most recently, Courvoisier et al. (2008) have combined this approach with latent state-trait analysis; the latter method allows one to estimate variance due to stable traits, occasion-specific states, and error (Steyer et al. 1999). The result is a single analytic method to estimate variance due to trait, method, state, and error. Among the possibilities offered by this approach is that one can investigate the degree to which method effects are stable or variable over time.

We wish to emphasize three points concerning these advances in methods for the empirical evaluation of construct validity. First, the concern that MTMM data could not successfully be analyzed using CFA/SEM approaches is no longer valid. There are now analytic tools that have proven successful (Eid et al. 2003). Second, statistical tools are available that enable one to quantitatively estimate multiple sources of variance that are important to the construct validation enterprise (Eid et al. 2003, Marsh & Grayson 1995). One need not guess at the degree to which method variance is present, or the degree to which it is common across traits, or the degree to which it is stable: One can investigate these sources of variance directly. Third, these analytic techniques are increasingly accessible to researchers (see Kline 2005 for a useful introduction to SEM). Clinical researchers have a validity concern beyond successful demonstration of convergent and discriminant validity. Success at the level of MTMM validity does not assure the measured traits have utility. Typically, one also needs to investigate whether the traits enhance prediction of some criterion of clinical importance.

To this end, clinical researchers can rely on a classic contribution by Hammond et al. (1986). They offered a creative, integrative analytic approach for combining the results of MTMM designs with the evaluation of differential prediction of external criteria. In the best tradition of applying basic science advances to practical prediction, their design integrated the convergent/discriminant validity perspective of Campbell & Fiske (1959) with Brunswik's (1952, 1956) emphasis on representative design in research, which in part concerned the need to conduct investigations that yield findings one can apply to practical problems. They presented the concept of a performance validity matrix, which adds criterion variables for each trait to the MTMM design. By adding clinical outcome variables to one's MTMM design, one can provide evidence of convergent validity, discriminant validity, and differential clinical prediction in a single study.

Such analyses are critical clinically because this sophisticated treatment of validity is likely to improve the usefulness of measures for clinicians. For many measures, validation research that considers practical prediction improves measures' "three Ps": predicting important criteria; prescribing treatments, and understanding the processes underlying personality and psychopathology (Youngstrom 2008), thereby improving clinical assessment. Such practical efforts in assessment must rely on observed scores, confounded as they may be with method variance. Construct validity research provides the clinician with an appreciation of the many

factors entering into an observed score and, thus, appreciation of the mix of construct-relevant variance, reliable construct-irrelevant variance, and method variance in any score (see Richters 1992).

CONCLUSION

The term "construct validation" refers to the process of simultaneously validating measures of psychological constructs and the theories of which the constructs are a part. The study of the construct validation process is ongoing. It rests on core principles identified 50 years ago (Campbell & Fiske 1959, Cronbach & Meehl 1955, Loevinger 1957, MacCorquodale & Meehl 1948); those principles remain central to theory testing today. It is also true that our understanding of construct validation has evolved over these 50 years.

In this review, we emphasized five ways in which this is true. First, advances in philosophy of science have helped clarify the ongoing, indeterminate nature of the construct validation process. This quality of theory testing represents a strength to the scientific method because it reflects the continuing process of critical evaluation of all aspects of theory and measurement. Second, theoreticians now emphasize the pursuit of informative theory tests in order to avoid weak, ad hoc theory tests in the absence of fully specified theories. Third, the need to validate clinical laboratory tasks, by investigating the degree to which responses on a task do reflect the influence of the target construct of interest, is becoming increasingly appreciated. Fourth, the lack of clarity that follows the use of a single score to represent multidimensional constructs has been described; researchers are increasingly relying on unidimensional measures to improve the validity of their theory tests. And fifth, important advances in the means to evaluate validity evidence empirically have been described; researchers have important new statistical tools at their disposal.

In sum, there are exciting new developments in the study of how to validate theories and their accompanying measures. These advances promise important improvements in measure and theory validation. As researchers fully incorporate sound construct validation theory in their methods, the rate of progress in clinical psychology research will continue to increase.

SUMMARY POINTS

1. The core perspective on construct validation can be understood by considering four classic papers published 50 or more years ago: Campbell & Fiske (1959), Cronbach & Meehl (1955), Loevinger (1957), and MacCorquodale & Meehl (1948).
2. Measures of psychological constructs are validated by testing whether they relate to measures of other constructs as specified by theory. Each test of relations between measures reflects on the validity of both the measures and the theory driving the test. Construct validation concerns the simultaneous process of measure and theory validation.
3. Current nonjustificationist philosophy of science indicates that no single experiment fully proves or disproves a theory. Instead, evidence for the validity of a measure and of a theory accrues over time and reflects an analysis of the evidence for and against theory claims.
4. Validation theorists are promoting an increased reliance on theory in clinical research because, it is argued, advances in knowledge are facilitated by theory-driven research.
5. Experimental psychopathology researchers using laboratory tasks should seek evidence that variation in performance on a task reflects variation in the psychological processes of interest more than does reliably measured but theoretically irrelevant constructs.

6. Validation efforts benefit when single scores reflect variation on a single dimension of psychological functioning. If a single score reflects multiple dimensions, variation among individuals on that score lacks unambiguous meaning.
7. New statistical tools are available for the quantitative investigation of many aspects of construct validity, including the assessment of convergent and discriminant validity. These tools are increasingly accessible to researchers.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Lee Anna Clark and Eric Youngstrom for their most helpful comments and suggestions, and Jill White for her excellent work in preparing the manuscript. Portions of this work were supported by NIAAA grant 1 RO 1 AA 016166 to Gregory T. Smith.

LITERATURE CITED

Am. Psychol. Assoc. 1954. Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull. Suppl.* 51:1–38

Anastasi A. 1950. The concept of validity in the interpretation of test scores. *Educ. Psychol. Meas.* 10:67–78

Barch DM. 2005. The cognitive neuroscience of schizophrenia. *Annu. Rev. Clin. Psychol.* 1:321–53

Barch DM, Carter CS, MacDonald AW, Braver TS, Cohen JD. 2003. Context-processing deficits in schizophrenia: diagnostic specificity, 4-week course, and relationships to clinical symptoms. *J. Abnorm. Psychol.* 112(1):132–43

Barrett P. 2005. What if there were no psychometrics? Constructs, complexity, and measurement. *J. Personal. Assess.* 85:134–40

Bartley WW III. 1962. *The Retreat to Commitment*. New York: Knopf

Bartley WW III. 1987. Philosophy of biology versus philosophy of physics. In *Evolutionary Epistemology, Rationality, and the Sociology of Knowledge*, ed. G Radnitzky, WW Bartley III, pp. 7–46. La Salle, IL: Open Court

Bechtoldt HP. 1951. Selection. In *Handbook of Experimental Psychology*, ed. SS Stevens, pp. 1237–66. Oxford, UK: Wiley

Blumberg AE, Feigl H. 1931. Logical positivism. *J. Philos.* 28:281–96

Borsboom D. 2006. The attack of the psychometricians. *Psychometrika* 71:425–40

Borsboom D, Mellenbergh GJ, van Heerden J. 2003. The theoretical status of latent variables. *Psychol. Rev.* 110:203–19

Borsboom D, Mellenbergh GJ, van Heerden J. 2004. The concept of validity. *Psychol. Rev.* 111:1061–71

Overviews philosophical, methodological, and substantive issues in validity theory.

Brinkley CA, Newman JP, Widiger TA, Lynam DR. 2004. Two approaches to parsing the heterogeneity of psychopathy. *Clin. Psychol. Sci. Pract.* 11:69–94

Browne MW. 1984. The decomposition of multitrait-multimethod matrices. *Br. J. Math. Stat. Psychol.* 37:1–21

Brunswik E. 1952. The conceptual framework of psychology. In *International Encyclopedia of Unified Science*, ed. O Neurath, R Carnap, C Morris, 1(10):1–102. Chicago: Univ. Chicago Press

Brunswik E. 1956. *Perception and the Representative Design of Psychological Experiments*. Berkeley: Univ. Calif. Press. 2nd ed.

Bryant FB. 2000. Assessing the validity of measurement. In *Reading and Understanding MORE Multivariate Statistics*, ed. G Laurence, PR Yarnold, pp. 99–146. Washington, DC: Am. Psychol. Assoc.

Butcher JN. 1990. *Use of the MMPI-2 in Treatment Planning*. New York: Oxford Univ. Press
Butcher JN. 1995. *Clinical Personality Assessment: Practical Approaches*. New York: Oxford Univ. Press
Butcher JN, ed. 2006. *A Practitioner's Guide*. Washington, DC: Am. Psychol. Assoc.
Butcher JN. 2002. Assessing pilots with "the wrong stuff": a call for research on emotional health factors in commercial aviators. *Int. J. Select. Assess.* 10:1–17
Campbell DT. 1987. Evolutionary epistemology. In *Evolutionary Epistemology, Epistemology, Rationality, and the Sociology of Knowledge*, ed. G Radnitzky, WW Bartley III, pp. 47–89. La Salle, IL: Open Court
Campbell DT. 1990. The Meehlian Corroboration-Verisimilitude theory of science. *Psychol. Inq.* 1(2):142–47
Campbell DT, Fiske DW. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56(2):81–105
Campbell DT, O'Connell EJ. 1967. Method factors in multitrait-multimethod matrices: multiplicative rather than additive? *Multivar. Behav. Res.* 2:409–26
Campbell DT, O'Connell EJ. 1982. Methods as diluting trait relationships rather than adding irrelevant systematic variance. In *New Directions for Methodology of Social and Behavioral Science: Forms of Validity in Research*, ed. D Brinberg, L Kidder, pp. 93–111. San Francisco, CA: Jossey-Bass
Carpenter PA, Just MA, Shell P. 1990. What one intelligence test measures: a theoretical account of processing in the Raven's Progressive Matrices Test. *Psychol. Rev.* 97:404–31
Caspi A. 1993. Why maladaptive behaviors persist: sources of continuity and change across the life course. In *Studying Lives Through Time: Personality and Development*, ed. DC Funder, RD Parke, CA Tomilinson-Keasey, K Widaman, pp. 343–76. Washington, DC: Am. Psychol. Assoc.
Caspi A, Roberts BW. 2001. Personality development across the life course: the argument for change and continuity. *Psychol. Inq.* 12:49–66
Clark LA. 2006. When a psychometric advance falls in the forest. *Psychometrika* 71:447–50
Clark LA, Watson D. 1995. Constructing validity: basic issues in objective scale development. *Psychol. Assess.* 7:309–19
Cohen JD, Barch DM, Carter C, Servan-Schreiber D. 1999. Context-processing deficits in schizophrenia: converging evidence from three theoretically motivated cognitive tasks. *J. Abnorm. Psychol.* 108(1):120–33
Cook TD, Campbell DT. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin
Cornblatt BA, Keilp JG. 1994. Impaired attention, genetics, and the pathophysiology of schizophrenia. *Schizophr. Bull.* 20:31–46
Costa PT Jr, McCrae RR. 1992. *Revised NEO Personality Inventory (NEO-PI–R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Odessa, FL: Psychol. Assess. Resourc.
Costa PT Jr, McCrae RR. 1995. Domains and facets: hierarchical personality assessment using the revised NEO personality inventory. *J. Personal. Assess.* 64:21–50
Courvoisier DS, Nussbeck FW, Eid M, Geiser C, Cole DA. 2008. Analyzing the convergent and discriminant validity of states and traits: development and applications of multimethod latent state-trait models. *Psychol. Assess.* 20(3):270–80
Cronbach LJ. 1988. Five perspectives on validation argument. In *Test Validity*, ed. H Wainer, H Braun, pp. 3–17. Hillsdale, NJ: Erlbaum
Cronbach LJ. 1995. Giving method variance its due. In *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske*, ed. PE Shrout, ST Fiske, pp. 145–60. Hillsdale, NJ: Erlbaum
Cronbach LJ, Gleser GC. 1965. *Psychological Test and Personnel Decisions*. Urbana: Univ. Ill. Press
Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. *Psychol. Bull.* 52:281–302
Cudeck R. 1988. Multiplicative models and MTMM matrices. *J. Educ. Stat.* 13:131–47
Cureton EE. 1950. Validity. In *Educational Measurement*, ed. EF Lingquist, pp. 621–94. Washington, DC: Am. Council Educ.
Depue RA, Collins PF. 1999. Neurobiology of the structure of personality: dopamine, facilitation of incentive motivation, and extraversion. *Behav. Brain Sci.* 22:491–517; discussion 518–69
Derksen J, Gerits L, Verbruggen A. 2003. *MMPI-2 profiles of nurses caring for people with severe behavior problems*. Presented at MMPI-2 Conf. Recent Develop. Use of MMPI-2 and MMPI-A, 38th, Minneapolis, Minn.

Provides a classic introduction of multitrait, multimethod methodology.

Provides an insightful review of the extent of progress in implementing sound validation procedures.

One of the defining articles on validation theory in psychology, it played a central role in introducing the concept of construct validity to psychology research.

Dimidjian S, Hollon SD, Dobson KS, Schmaling KB, Kohlenberg RJ, et al. 2006. Randomized trial of behavioral activation, cognitive therapy and antidepressant medication in the acute treatment of adults with major depression. *J. Consult. Clin. Psychol.* 74:658–70

Duhem P. 1914. *The Aim and Structure of Physical Theory*. Transl. P Weiner, 1991. Princeton, NJ: Princeton Univ. Press (from French)

Edwards JR. 2001. Multidimensional constructs in organizational behavior research: an integrative analytical framework. *Organ. Res. Methods* 4:144–92

Eid M, Lischetzke T, Nussbeck FW, Trierweiler LI. 2003. Separating trait effects from trait-specific method effects in multitrait-multimethod models: a multiple-indicator CT-C(M-1) model. *Psychol. Methods* 8:38–60

Embretson SE. 1998. A cognitive design system approach for generating valid tests: approaches to abstract reasoning. *Psychol. Methods* 3:300–96

Engle RW, Tuholski SW, Laughlin JE, Conway AR. 1999. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *J. Exp. Psychol.: Gen.* 128(3):309–31

Feyerabend P. 1970. Against method. In *Minnesota Studies on the Philosophy of Science. Vol. IV: Analyses of Theories and Methods of Physics and Psychology*, ed. M Radner, S Winokur, pp. 17–130. Minneapolis: Univ. Minn. Press

Flemming EG, Flemming CW. 1929. The validity of the Matthews' revision of the Woodworth personal data questionnaire. *J. Abnorm. Soc. Psychol.* 23:500–6

Fossati A, Citterio A, Grazioli F, Borroni S, Carretta I, et al. 2005. Taxonic structure of schizotypal personality disorder: a multi-instrument, multi-sample study based on mixture models. *Psychiatr. Res.* 137:71–85

Analyzes the critical role of multiple methods for the definition of a construct.

Garner WR, Hake HW, Eriksen CW. 1956. Operationism and the concept of perception. *Psychol. Rev.* 63(3):149–59

Garrett HE, Schneck MR. 1928. A study of the discriminate value of the Woodworth Personal Data Sheet. *J. Gen. Psychol.* 1:459–71

Gass CS. 2002. Personality assessment of neurologically impaired patients. In *Clinical Personality Assessment: Practical Approaches*, ed. J Butcher, pp. 208–44. New York: Oxford Univ. Press. 2nd ed.

Goldberg LR. 1993. The structure of personality traits: vertical and horizontal aspects. In *Studying Lives Through Time: Personality and Development*, ed. DC Funder, RD Parke, C Tomlinson-Keasey, K Widaman, pp. 169–88. Washington, DC: Am. Psychol. Assoc.

Goldman DS. 2007. Legal construct validation: expanding empirical legal scholarship to unobservable concepts. *bepress Legal Ser.* Work. pap. 1715. **http://law.bepress.com/expresso/eps/1715**

Gough HG. 1996. *CPI Manual*. Palo Alto, CA: Consult. Psychol. Press

Greene RL. 2006. Use of the MMPI-2 in outpatient mental health settings. See Butcher 2006, pp. 253–72

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychol. Assess.* 12:19–30

Hacking I. 1999. *The Social Construction of What*? Cambridge, MA: Harvard Univ. Press

Hammond KR, Hamm R, Grassia J. 1986. Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychol. Bull.* 100:257–69

Hough LM, Schneider RJ. 1995. Personality traits, taxonomies, and applications in organizations. In *Individuals and Behavior in Organizations*, ed. KR Murphy, pp. 31–88. San Francisco, CA: Jossey-Bass

Jang KL, Livesley WJ, Taylor S, Stein MB, Moon EC. 2004. Heritability of individual depressive symptoms. *J. Affect. Disord.* 80:125–33

Kane MT. 2001. Current concerns in validity theory. *J. Educ. Meas.* 38:319–42

Kane MT. 2006. In praise of pluralism: a comment on Borsboom. *Psychometrika* 71:441–45

Kenny DA, Kashy DA. 1992. Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychol. Bull.* 112:165–72

King D, Leskin G, King L, Weathers F. 1998. Confirmatory factor analysis of the clinical administered PTSD scale: evidence for the dimensionality of posttraumatic stress disorder. *Psychol. Assess.* 10:90–96

Kline RB. 2005. *Principles and Practice of Structural Equation Modeling*. New York: Guilford

Knight RA, Silverstein SM. 2001. A process-oriented approach for averting confounds resulting from general performance deficiencies in schizophrenia. *J. Abnorm. Psychol.* 110:15–30

Krueger RF, McGue M, Iacono WG. 2001. The higher-order structure of common DSM mental disorders: internalization, externalization, and their connections to personality. *Personal. Individ. Differ.* 30:1245–59
Kuhn TS. 1970. *The Structure of Scientific Revolutions*. Chicago: Univ. Chicago Press
Kusch M. 2002. Metaphysical déjà vu: Hacking and Latour on science studies and metaphysics. *Stud. Hist. Philos. Sci.* 33:639–47
Laird DA. 1925. A mental hygiene and vocational test. *J. Educ. Psychol.* 16:419–22
Lakatos I. 1968. Criticism and the methodology of scientific research programs. *P. Aristotelian Soc.* 69:149–86
Lakatos I. 1999. Lectures on scientific method. In *For and Against Method*, ed. I Lakatos, P Feyerabend, pp. 19–112. Chicago: Univ. Chicago Press
Landy FJ. 1986. Stamp collecting versus science: validation as hypothesis testing. *Am. Psychol.* 41:1183–92
Latour B. 1999. *Essays on the Reality of Science Studies*. Cambridge, MA: Harvard Univ. Press
Linehan MM. 1993. *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. New York: Guilford
Linehan MM, Heard HL, Armstrong HE. 1993. Naturalistic follow-up of a behavioral treatment for chronically parasuicidal borderline patients. *Arch. Gen. Psychiatry* 50:971–74
Loevinger J. 1954. The attenuation paradox in test theory. *Psychol. Bull.* 51:493–504
Loevinger J. 1957. Objective tests as instruments of psychological theory. *Psychol. Rep. Monogr. Suppl.* 3:635–94
Lynam DR, Widiger TA. 2007. Using a general model of personality to identify the basic elements of psychopathy. *J. Personal. Disord.* 21:160–78
MacCorquodale K, Meehl PE. 1948. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.* 55(2):95–107
MacDonald AW, Goghari VM, Hicks BM, Flory JD, Carter CS, Manuck SB. 2005. A convergent-divergent approach to context processing, general intellectual functioning, and the genetic liability to schizophrenia. *Neuropsychology* 19(6):814–21
Maraun MD, Peters J. 2005. What does it mean that an issue is conceptual in nature? *J. Personal. Assess.* 85:128–33
Marsh HW, Grayson D. 1995. Latent variable models of multitrait-multimethod data. In *Structural Equation Modeling: Concepts, Issues, and Application*, ed. RH Hoyle, pp. 177–98. London: Sage
McCrae R, Zonderman A, Costa P, Bond M, Paunonen S. 1996. Evaluating replicability of factors in the revised NEO Personality Inventory: confirmatory factor analysis versus Procrustes rotation. *J. Personal. Soc. Psychol.* 70:552–66
McGrath RE. 2005. Conceptual complexity and construct validity. *J. Personal. Assess.* 85:112–24
Meehl PE. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46:806–34
Meehl PE. 1990. Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant it. *Psychol. Inq.* 1(2):108–41
Meehl PE. 1992. Factors and taxa, traits and types, differences of degree and differences in kind. *J. Personal.* 60:117–73
Megargee EI. 2006. Use of the MMPI-2 in correctional settings. See Butcher 2006, pp. 327–60
Megargee EI. 2008. The California Psychological Inventory. In *Oxford Handbook of Personality Assessment*, ed. JN Butcher. New York: Oxford Univ. Press. In press
Messick S. 1989. Validity. In *Educational Measurement*, ed. RL Linn, pp. 13–103. New York: Am. Counc. Educ./Macmillan. 3rd ed.
Messick S. 1995. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50(9):741–49
Michell J. 2000. Normal science, pathological science and psychometrics. *Theor. Psychol.* 10:639–67
Michell J. 2001. Teaching and misteaching measurement in psychology. *Aust. Psychol.* 36:211–17
Millon T, Davis RD, Millon CM, Wenger AW, Van Zuilen MH, et al. 1996. *Disorders of Personality. DSM–IV and Beyond.* New York: Wiley
Morey L. 2002. Measuring personality and psychopathology. In *Handbook of Psychology. Vol. 2: Research Methods in Psychology*, ed. J Schinka, W Velicer, I Weiner, pp. 377–406. Hoboken, NJ: Wiley
Nichols DS, Crowhurst B. 2006. Use of the MMPI-2 in inpatient mental health settings. See Butcher 2006, pp. 195–252

Integratively presents construct validity as subsuming specific forms of validation tests.

Promotes the philosophical legitimacy of hypothetical constructs.

Emphasizes that validity refers to interpretations of test scores.

Nuechterlein KH. 1991. Vigilance in schizophrenia and related disorders. In *Handbook of Schizophrenia, Neuropsychology, Psychophysiology and Information Processing*, ed. SR Steinhauer, JH Gruzelier, J Zubin, pp. 397–433. Amsterdam: Elsevier Sci.

Nunnally JC, Bernstein IH. 1994. *Psychometric Theory*. New York: McGraw-Hill

Ozer D. 1989. Construct validity in personality assessment. In *Personality Psychology: Recent Trends and Emerging Directions*, ed. DM Buss, N Cantor, pp. 224–34. New York: Springer-Verlag

Paunonen SV. 1998. Hierarchical organization of personality and prediction of behavior. *J. Personal. Soc. Psychol.* 74:538–56

Paunonen SV, Ashton MC. 2001. Big Five factors and facets and the prediction of behavior. *J. Personal. Soc. Psychol.* 81:524–39

Perry JN, Miller KB, Klump K. 2006. Treatment planning with the MMPI-2. See Butcher 2006, pp. 143–64

Phillips WA, Silverstein SM. 2003. Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia: a physiological, computational, and psychological perspective. *Behav. Brain Sci.* 26:65–138

Posner MI. 1978. *Chronometric Explorations of Mind*. Hillsdale, NJ: Erlbaum

Reichardt CS, Coleman SC. 1995. The criteria for convergent and discriminant validity in a multitrait-multimethod matrix. *Multivar. Behav. Res.* 30:513–38

Richters JE. 1992. Depressed mothers as informants about their children: a critical review of the evidence for distortion. *Psychol. Bull.* 112:485–99

Roberts BW, DelVecchio WF. 2000. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. *Psychol. Bull.* 126:3–25

Schneider RJ, Hough LM, Dunnette MD. 1996. Broadsided by broad traits: how to sink science in five dimensions or less. *J. Organ. Behav.* 17:639–55

Sechrest L, Davis M, Stickle T, McKnight P. 2000. Understanding "method" variance. In *Research Design: Donald Campbell's Legacy*, ed. L Bickman, 2:63–87. Thousand Oaks, CA: Sage

Serretti A, Olgiati P. 2004. Dimensions of major psychoses: a confirmatory factor analysis of six competing models. *Psychiatr. Res.* 127:101–9

Silverstein SM. 2008. Measuring specific, rather than generalized, cognitive deficits and maximizing between-group effect size in studies of cognition and cognitive change. *Schizophr. Bull.* 34:645–55

Simms L, Watson D, Doebbeling B. 2002. Confirmatory factor analyses of posttraumatic stress symptoms in deployed and nondeployed veterans of the Gulf War. *J. Abnorm. Psychol.* 111:637–47

Smith GT. 2005. On construct validity: issues of method and measurement. *Psychol. Assess.* 17:396–408

Smith GT, Combs J. 2008. Issues of construct validity in psychological diagnoses. In *Contemporary Directions in Psychopathology: Toward the DSM-V and ICD-11*, ed. T Millon, RF Krueger, E Simonsen. New York: Guilford. In press

Smith GT, Fischer S, Cyders MA, Annus AM, Spillane NS, McCarthy DM. 2007. On the validity and utility of discriminating among impulsivity-like traits. *Assessment* 14(2):155–70

Smith GT, Fischer S, Fister SM. 2003. Incremental validity principles in test construction. *Psychol. Assess.* 15(4):467–77

Smith GT, McCarthy DM. 1995. Methodological considerations in the refinement of clinical assessment instruments. *Psychol. Assess.* 7:300–8

Steyer R, Schmitt M, Eid M. 1999. Latent state-trait theory and research in personality and individual differences. *Eur. J. Personal.* 13:389–408

Strauss ME. 2001. Demonstrating specific cognitive deficits: a psychometric perspective. *J. Abnorm. Psychol.* 110:6–14

Strauss ME, Summerfelt AT. 1994. Response to Serper and Harvey's "On integrating cognitive psychology and neuropsychology in schizophrenia research." *Schizophr. Bull.* 20:13–21

Swets JA, Dawes RM, Monahan J. 2000. Psychological science can improve diagnostic decisions. *Psychol. Sci. Public Interest* 1:1–26

Uhlhaas PJ, Silverstein SM. 2005. Perceptual organization in schizophrenia spectrum disorders: empirical research and theoretical implications. *Psychol. Bull.* 131(4):618–32

Watson D, Clark LA, Carey G. 1988a. Positive and negative affect and their relation to anxiety and depressive disorders. *J. Abnorm. Psychol.* 97:346–53

Reviews recent advances in construct validation theory, including increased precision in differentiating among clinical constructs, and ongoing efforts to improve the construct validation process.

Watson D, Clark LA, Tellegen A. 1988b. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Personal. Soc. Psychol.* 54:1063–70

Wechsler D. 1997. *Manual for the Wechsler Adult Intelligence Scale–3rd Edition*. New York: Psychol. Corp.

Weimer WB. 1979. *Notes on the Methodology of Scientific Research*. Hillsdale, NJ: Erlbaum

Whitely SE. 1983. Construct validity: construct representation versus nomothetic span. *Psychol. Bull.* 93:179–97

Whiteside SP, Lynam DR. 2001. The five factor model and impulsivity: using a structural model of personality to understand impulsivity. *Personal. Individ. Differ.* 30:669–89

Widaman KF. 1985. Hierarchically nested covariance structure models for multitrait-multimethod data. *Appl. Psychol. Meas.* 9:1–26

Widiger TA, Simonsen E. 2005. Alternative dimensional models of personality disorder: finding a common ground. *J. Personal. Disord.* 19:110–30

Widiger TA, Simonsen E, Krueger R, Livesley WJ, Verheul R. 2005. Personality disorder research agenda for the DSM-V. *J. Personal. Disord.* 19:317–40

Widiger TA, Trull TJ. 2007. Plate tectonics in the classification of personality disorder: shifting to a dimensional model. *Am. Psychol.* 62:71–83

Wothke W. 1995. Covariance components analysis of the multitrait-multimethod matrix. In *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske*, ed. PE Shrout, ST Fiske, pp. 125–44. Hillside, NJ: Erlbaum

Youngstrom E. 2008. Evidence-based strategies for the assessment of developmental psychopathology: measuring prediction, prescription, and process. In *Psychopathology: History, Diagnosis and Empirical Foundations*, ed. WE Craighead, DJ Miklowitz, LW Craighead, pp. 34–77. Hoboken, NJ: Wiley

The seminal analysis of the distinction between nomothetic span and construct representation.

Annual Review of
Clinical Psychology
Volume 5, 2009

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Clinical Psychology* articles may be found at http://clinpsy.annualreviews.org