Psychological Assessment
2011, Vol. 23, No. 2, 532–544

1040-3590/11/$12.00 DOI: 10.1037/a0022402

Toward a Process-Focused Model of Test Score Validity: Improving Psychological Assessment in Science and Practice

Robert F. Bornstein
Adelphi University

Although definitions of validity have evolved considerably since L. J. Cronbach and P. E. Meehl's classic (1955) review, contemporary validity research continues to emphasize correlational analyses assessing predictor–criterion relationships, with most outcome criteria being self-reports. The present article describes an alternative way of operationalizing validity—the process-focused (PF) model. The PF model conceptualizes validity as the degree to which respondents can be shown to engage in a predictable set of psychological processes during testing, with those processes dictated a priori by the nature of the instrument(s) used and the context in which testing takes place. In contrast to the traditional approach wherein correlational methods are used to quantify the relationship between test score and criterion, the PF model uses experimental methods to manipulate variables that moderate test score–criterion relationships, enabling researchers to draw more definitive conclusions regarding the impact of underlying psychological processes on test scores. By complementing outcome-based validity assessment with a process-driven approach, researchers will not only improve psychology's assessment procedures but also enhance their understanding of test bias and test score misuse by illuminating the intra- and interpersonal factors that lead to differential performance (and differential prediction) in different groups.

Keywords: validity, construct validity, psychological assessment, psychometrics, test bias

If there is a single challenge that characterizes all of psychology's diverse subfields, that challenge is assessment. Psychologists measure things. These "things" take many forms, including observable behaviors and hidden mental states; dyadic interactions and intergroup dynamics; changes in traits, symptoms, skills, and abilities over time; and a broad array of neurophysiological and neurochemical processes (along with their associated behaviors and mental activities).

Given psychologists' near universal reliance on assessment, it is not surprising that researchers have devoted considerable time and effort to maximizing test score validity—to ensuring that researchers' assessment tools measure what we think they do. Maximizing test score validity is not merely an academic exercise, but one that goes to the heart of psychological science and practice, with widespread social implications. The availability of measures that yield scores with strong validity evidence enables psychologists to enhance the scientific rigor of their research, make accurate decisions in applied settings, and use and interpret test results fairly, in unbiased ways. To be sure, the existence of assessment tools that yield well-validated scores does not ensure scientific rigor and accurate, unbiased decision making, but the absence of such tools guarantees that neither of these things will occur. Thus, continued efforts to increase our understanding of test score validity and improve our validation methods will benefit the science and practice of psychology in myriad ways.

The present article contributes to that effort by describing an approach to validity—the process-focused (PF) Model—that differs markedly from the traditional perspective. In contrast to the traditional approach wherein the heart of validity lies in outcome—in the relation of predictor scores to some criterion measure—the PF model conceptualizes validity as the degree to which respondents can be shown to engage in a predictable set of psychological processes during assessment, with those processes dictated a priori by the nature of the instrument(s) used, and context in which testing takes place. The PF model differs from traditional validity assessment not only with respect to how validity is conceptualized but also with respect to empirical emphasis: In contrast to the traditional approach wherein correlational methods are used to quantify the relationship between test score and criterion, the PF model uses experimental methods to manipulate variables that moderate test score–criterion relationships, enabling researchers to draw more definitive conclusions regarding the impact of underlying processes (e.g., autobiographical memory search in response to a self-report questionnaire item) and moderating variables (e.g., motivation, mood state) on test scores.

In short, the PF model shifts the emphasis of validity theory and research from outcome to process, and from correlation to exper-

This material is based on work supported by National Science Foundation Grant 0822010. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. I thank Joseph Masling, Carolyn Morf, and Kathleen Slaney for their helpful comments on earlier versions of this paper and Julia Biars and Alexandra Rosen for help in collecting and coding studies included in Table 2.

Correspondence concerning this article should be addressed to Robert F. Bornstein, Derner Institute of Advanced Psychological Studies, Adelphi University, 212 Blodgett Hall, Garden City, NY 11530. E-mail: bornstein@adelphi.edu

imentation. By complementing traditional validity assessment with a process-driven approach, we will not only improve psychology's assessment procedures but also enhance researchers' understanding of test bias and test score misuse by illuminating the underlying intra- and interpersonal dynamics that lead to differential performance (and differential prediction) in different groups.

I begin by reviewing the traditional concept of validity and its limitations, and the evolution of validity theory and research during the past several decades. I then outline an alternative process-focused model of validity. I discuss how data from process-focused and outcome-focused studies may be combined and conclude by elaborating research, practice, and social policy implications of an integrated perspective.

The Traditional Conceptualization of Test Score Validity

Although psychology has a long history of using standardized assessment instruments to quantify aptitude, attitude, achievement, personality, and psychopathology, contemporary validity theory and research began in earnest in the mid-20th century, with the publication of Cronbach and Meehl's (1955) "Construct Validity in Psychological Tests," and 4 years later, Campbell and Fiske's (1959) "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." Most theoretical analyses and empirical studies of validity during the past 50 years have taken as their starting point ideas and assumptions outlined in these two seminal papers.[1]

Following the logic of Cronbach and Meehl (1955) and Campbell and Fiske (1959), validity has traditionally been operationalized as a statistic, the *validity coefficient* (usually expressed as *r*), that reflects the magnitude of the relationship between a predictor (the test score) and some criterion (an outcome measure). A wide variety of criteria are predicted by psychological tests, some overt and readily observable, others hidden and only detectable indirectly. When an observable criterion (e.g., suicide attempts) is assessed, the validity coefficient is said to be an index of *criterion validity*; when an unobservable construct (e.g., suicidal ideation) is assessed, the validity coefficient is an index of *construct validity*. Construct validity is in turn divided into *convergent validity* (the degree to which a test score is associated with some theoretically related variable), and *discriminant validity* (the degree to which a test score is unrelated—or minimally related—to a theoretically unrelated variable). Criterion validity can be operationalized in terms of *concurrent validity* (when the test score is used to assess some outcome in the here-and-now), and *predictive validity* (when the test score is used to predict some outcome in the future), although as psychometricians have noted, the point in time at which concurrent validity morphs into predictive validity is difficult to specify, and varies as a function of the criterion being assessed and purpose of the assessment.

There are also a number of validity-related variables that are not indices of validity in its strictest sense, but are nonetheless germane in the present context. For example, researchers often seek to quantify *internal validity* using analyses of individual test items and response patterns to derive estimates of internal consistency and factor structure. Though these coefficients do not address issues regarding the degree to which test scores predict external variables or outcomes, when internal reliability data conform to a priori expectations regarding item interrelationships, factors, and clusters, these data indirectly support the construct validity of scores derived from the measure. Test score reliability—especially retest reliability—also has implications for validity: When a test is designed to quantify a construct that is presumed to be stable over time (e.g., manual dexterity, narcissism), inadequate retest reliability is prima facie evidence of a problem with test score validity. Most psychometricians agree that face validity—test "obviousness"—is not a true index of validity.[2]

It is important to note that regardless of whether one is considering criterion, construct, convergent, or discriminant validity, traditional validity coefficients all have two things in common. First, they are all indices of strength of association, and represented as correlations of one kind or another. Second, they are all assessed observationally. Even when test scores and outcome measures are obtained in laboratory settings under highly controlled conditions, they invariably reflect the pairing of data collected during one period of observation (the testing) with data collected during a second period of observation (the comparison measure).

Although traditional correlational methods continue to be widely used in validity research, in recent years psychometricians have increasingly used confirmatory factor analysis (CFA)—a variant of structural equation model (SEM)—to examine hypothesized causal relations among variables and draw inferences regarding underlying process links (see Hershberger, 2003, for a historical review). SEM is particularly useful in enabling researchers to delineate latent variables—variables not assessed directly by the tests administered, but which emerge from a well-specified model when predicted patterns of variable intercorrelations are obtained (see Ullman, 2006). In certain instances, these latent variables represent underlying, unobservable psychological processes; to obtain more definitive evidence regarding underlying process links, latent variables that emerge from SEM analyses may then be explored via experimental studies wherein key parameters are manipulated (Bollen & Davis, 2009; Hagemann & Meyerhoff, 2008; Schumacker & Lomax, 2004). As Schumacker and Lomax (2004) noted,

> In structural equation modeling, the amount of influence rather than cause-and-effect relationship is assumed and interpreted by direct, indirect, and total effects among variables . . . . Model testing involves the use of manipulative variables, which, when changed, affect the

[1] Other seminal early papers on construct validity were MacCorquodale and Meehl's (1948) discussion of the epistemological challenges involved in validating scores derived from measures of hypothetical constructs and Loevinger's (1957) discussion of construct validation as one component of psychologists' broader efforts to develop and refine theoretical concepts.

[2] In this context, it is important to distinguish the narrow use of the terms *internal* and *external validity*, as these terms apply to test scores from the more general use of these terms by Campbell and Stanley (1963), who discussed various threats to the integrity of psychological assessments and experimental designs (see Slaney & Maraun, 2008).

model outcome values, and whose effects can hence be assessed." (p. 56)[3]

Refinements of the Traditional View

SEM and CFA have had a profound influence on validity research in recent years (see Hershberger, 2003; Tomarken & Waller, 2005; Ullman, 2006). Beyond these innovative statistical techniques, three substantive conceptual refinements of the traditional view of validity have emerged; each extends this view in an important way.

Construct Representation

Embretson (1983) distinguished the long-standing goal of construct validity, the weaving of a "nomological net" of relationships between test score and an array of theoretically related variables (which she termed "nomothetic span") from a complementary goal that she termed "construct representation": efforts to identify the theoretical mechanisms that underlie item responses. Drawing primarily from research on cognitive modeling, Embretson (1983, 1994; Embretson & Gorin, 2001) advocated the use of direct observation of testees, path analysis, posttest interview data, and other external indices to illuminate the processes in which people engage while completing psychological tests. Since Embretson's introduction of the concept of construct representation, numerous researchers have used these techniques to deconstruct the cognitive processes that occur when respondents engage items from various measures of aptitude, intelligence, and mental ability, adding expert ratings of item content and causal modeling techniques as additional methods for evaluating construct representation (see Cramer, Waldrop, van der Maas, & Borsboom, 2010; Kane, 2001; Mislevy, 2007).

Attribute Variation

Noting that measures of test score–criterion association provide limited information regarding the degree to which a test actually measures the variable it purports to assess, Borsboom, Mellenbergh, and van Heerden (2004) outlined an *attribute variation* approach, arguing that rigorous validity assessment requires demonstrating that changes in an attribute can be linked directly to changes in scores on a test designed to measure that attribute. Consistent with Embretson's (1983) construct representation view, Borsboom et al. (2004) suggested that "somewhere in the chain of events that occurs between item administration and item response, the measured attribute must play a causal role in determining what value the measurement's outcomes will take" (p. 1062). Emphasizing naturally occurring variations in traits and abilities rather than direct manipulation of underlying variables, Borsboom et al. cited latent class analyses that detect Piagetian developmental shifts in children's reasoning over time (e.g., Jansen & Van der Maas, 1997, 2002) as exemplars of the attribute variation approach (see also Strauss & Smith, 2009, for examples of the attribute variation approach in clinical assessment).

Consequential Validity

Originally described by Cronbach (1971), and elaborated extensively by Messick (1989, 1994, 1995), the concept of *consequential validity* represents a very different perspective on test score validation. According to this view, validity lies not only in the degree to which a test score is capable of predicting some theoretically related outcome but also in the degree to which that test score is actually used (and interpreted) in such a way as to yield valid data (see also Kane, 1992, for a related discussion). Thus, evidential (research-based) validity can be distinguished from consequential (impact-based) validity, the former representing a test's potential to provide accurate, useful, and unbiased information, and the latter representing the degree to which the test truly does yield accurate, useful, and unbiased assessment data in vivo. Inherent in the consequential validity framework is the assumption that an evidentially valid test score can provide consequentially valid data in certain contexts, but consequentially invalid data in others, depending on how the test score is interpreted. For example, intelligence test scores may be interpreted differently in two different schools, and psychopathology scores may be interpreted differently in two different clinics; in both situations, the test score in question might well yield consequentially valid information in one setting and consequentially invalid information in the other.

Validity Assessment in Theory and in Practice

As Jonson and Plake (1998) noted, a major trend in validity assessment since the mid-1950s has been a shift from conceptualizing validity in terms of discrete and separable subtypes (e.g., concurrent, predictive) to a more integrative approach wherein validity is conceptualized as a unitary concept (see Messick, 1995, for a detailed discussion of this issue). The notion that the overarching concept of construct validity can incorporate a broad spectrum of evidence was implied in Cronbach and Meehl's (1955) seminal paper, suggested more directly by Loevinger (1957), made explicit by Cronbach (1971) and Messick (1989), and is now the most widely accepted framework for unifying and integrating various aspects of test score validity (see Shepard, 1993; Slaney & Maraun, 2008).

Influenced by Messick's (1989, 1994, 1995) seminal conceptual analyses of test score validity (see also Cronbach, 1971), the most recent edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) describes validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests . . . . The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed test score interpretations." (p. 9). Consistent with most contemporary psychological assessment texts (e.g., McIntire

[3] Tomarken and Waller (2005) provided a particularly thorough and balanced review of the advantages and limitations of SEM, noting that although SEM is a powerful method for testing parameters of models that incorporate various combinations of observed and latent variables, it cannot provide definitive support for an hypothesized set of variable interrelations in the absence of external confirming evidence. Tomarken and Waller concluded that SEM "is arguably the most broadly applicable statistical procedure currently available", but went on to note that "SEM is not, however, a statistical magic bullet. It cannot be used to prove that a model is correct and it cannot compensate for a poorly designed study" (p. 56).

& Miller, 2000), the 1999 *Standards* still enumerates distinct types of validity evidence (see Table 1 for a summary of these categories), but also argues that distinctions among various types of validity evidence are less sharp than earlier frameworks had suggested and that multiple forms of converging evidence should be used to establish the validity of test scores within a particular context.

To the degree that psychologists have used the broad-based validation strategies described in the 1999 edition of the *Standards*, one would expect that during the past decade, psychology has moved beyond the traditional correlational-observational conceptualization of validity to a more integrative view. But have theoretical shifts in researchers' conceptualization of validity altered the practice of validity assessment in vivo? A review of present validity practices suggests that they have not.

Since publication of the 1999 *Standards*, there have been two major reviews of researchers' operationalization and assessment of test score validity. In the first, Hogan and Agnello (2004) surveyed 696 research reports from the American Psychological Association's *Directory of Unpublished Experimental Mental Measures* (Goldman & Mitchel, 2003), identifying the types of validity evidence reported for each measure. They found that for 87% of measures, the only validity evidence reported involved correlations between test scores and scores on other self-report scales. Only 5% of measures had been validated using behavioral outcome criteria (e.g., work performance, academic course grades). No entries reported validity evidence wherein experimental procedures were used to manipulate participants' response processes. Summarizing the implications of their findings, Hogan and Agnello (2004) concluded that despite recent definitional shifts, "the vast majority [of studies] reported correlations with other variables . . . . and little use was made of the numerous other types of validation approaches" (p. 802).

Similar results were obtained by Cizek, Rosenberg, and Koons (2008), who surveyed sources of validity evidence for all 283 tests reviewed in the 16th edition of the *Mental Measurements Yearbook* (MMY; Spies & Plake, 2005). Cizek et al. found that the vast majority of reports relied exclusively on correlational methods to evaluate test score validity, with only fivre of 283 entries (1.8%) assessing participant response processes. The proportion of MMY entries reporting process data was highest for developmental tests (5.9%), followed by behavioral measures (4.0%), achievement tests (3.7%), and tests of cognitive skills (1.5%). In every other test category (i.e., attitude, motor skill, personnel, personality, psychopathology, social, and vocational), the number of process-based validity studies was zero.

A review of recent empirical literature examining test score validity confirms the findings of Hogan and Agnello (2004) and Cizek et al. (2008). Table 2 summarizes these data, analyzing the measures and methods used by validity researchers in the five journals that have published the greatest number of validity studies during 2006–2008.[4] As Table 2 shows, despite shifts in the theoretical conceptualization of validity (e.g., Messick, 1989) and the recommendations of the 1999 *Standards*, the procedures used by assessment researchers to evaluate the validity of test scores remain largely unchanged. The vast majority of validity studies published in leading journals used correlational methods (91%), relying exclusively on self-report outcome measures (79%). Only 9% of studies in the five leading validity journals used experimental procedures.

To obtain a more complete picture of the methods used in correlational validity investigations in Table 2, the 442 studies in this category were classified into two subgroups: (a) studies in which SEM and CFA were used to examine patterns of variable interrelations and (b) studies that simply assessed the magnitude of association between predictor and criterion. Analysis revealed that 104 of 442 studies (24%) used CFA and/or SEM; the remaining 338 studies (76%) reported predictor-criterion correlational analyses. Proportions of studies in which SEM/CFA was used ranged from a low of 15% (*Journal of Personality Assessment*) to a high of 36% (*Educational and Psychological Measurement*), with *Psychological Assessment* (21%), *Assessment* (22%), and *Journal of Pychoeducational Assessment* (24%) falling between these extremes.

Table 1
Types of Validity Evidence in the 1999 Standards

Type of validity evidence	Typical validation procedure
Evidence based on test content	Logical analysis and expert ratings of item content, item-construct fit, item relevance, universe sampling, and criterion contamination
Evidence based on response processes	Interview-based and observational analyses of participants' responses to items or tasks; comparison of process differences across groups; studies of observer/interviewer decision processes
Evidence based on internal structure	Factor analyses, cluster analyses, item analyses, differential item functioning studies
Evidence based on relations to other variables	Concurrent and predictive validity, convergent and discriminant validity, validity generalization, criterion group differences, studies examining impact of interventions/manipulations on test scores, longitudinal studies
Evidence based on consequences of testing	Studies of expected/obtained benefits of testing; studies of unintended negative consequences

Note. A complete description of types of validity evidence and validation strategies is included in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Discussions of these validity categories are found in Goodwin and Leech (2003), Jonson and Plake (1998), and Messick (1995).

[4] These journals were identified via a PsycNFO search conducted in May 2009 using the keywords *Validity*, *Construct Validity*, *Criterion Validity*, and *Validation.* Interrater reliability in coding validity article characteristics was determined using procedures analogous to those of Cizek et al. (2008): All articles were coded by the author, and a second rater unaware of these initial codings independently recoded a sample of 100 articles (approximately 20% of the total). Agreement in coding was 98% for study design/method (correlational vs. experimental), and 93% for outcome/criterion (self-report vs. alternative measure).

Table 2
Validity Assessment Strategies, 2006–2008

Journal	Number of validity articles	Proportion of studies using correlational designs	Proportion of studies using experimental designs	Proportion of studies using self-report outcome measures	Proportion of studies using alternative outcome measures
Assessment	93	94%	6%	78%	22%
Educational & Psychological Measurement	93	92%	8%	91%	9%
Journal of Personality Assessment	131	89%	11%	75%	25%
Journal of Psycho educational Assessment	49	94%	6%	71%	29%
Psychological Assessment	120	91%	9%	76%	24%
Overall	486	91%	9%	79%	21%

Note. Validity articles include all articles that reported data regarding the construct validity of scores derived from a psychological test (even if these data were not explicitly identified as validity evidence by the authors). Only articles reporting original data were coded; literature reviews, meta-analyses, comments, and case studies were excluded. Studies using experimental designs used a manipulation with two or more conditions to contrast test responses in different groups (e.g., contrasting instructions, different pretest primes, treatment intervention vs. control/no treatment prior to test administration). Alternative outcome measures were any outcome measures not based on participant self-report (e.g., recording of participant behavior in laboratory or field, physiological measures, reports by knowledgeable informants, behaviors coded from archival/chart records).

Thus, the patterns in Table 2 confirm what earlier reviews had suggested: There is a substantial disconnect between the idealized descriptions of test score validity offered by psychometricians and the everyday practices of validity researchers in the laboratory, clinic, classroom, and field. Conceptual shifts notwithstanding, validity assessment in psychology remains much as it was 50 years ago.

One cannot ascertain from these data why the test score validation process has not changed appreciably in response to theoretical shifts. Continued reliance on traditional methods might be due in part to the fact that the definition of validity in the 1999 *Standards*—although potentially useful—is somewhat vague, without clear guidelines regarding operationalization, implementation, and integration of different forms of validity evidence. Slow progress in this area might also be due in part to the fact that most discussions of construct representation and the methods used to assess it have focused exclusively on tests of cognitive ability, not extending these principles to measures of personality, psychopathology, attitudes, interests, motives, and values (see Embretson, 1998; Embretson & Gorin, 2001; Kane, 2001; Mislevy, 2007). Our continued reliance on traditional validation methods might also reflect a more generalized reluctance among assessment researchers to move beyond well-established methods that—although flawed—are widely accepted in the scientific community.

Whatever the cause, recent theoretical refinements have had minimal impact on the validation efforts of psychometricians, and the difficulties that have long characterized the measurement of validity remain. Validity continues to be conceptualized primarily in terms of observation and correlation, and operationalized as a validity coefficient (or set of validity coefficients). Most psychological test scores yield validity coefficients in the small to moderate range (Meyer et al., 2001), predicting a modest amount of variance in outcome. Not only do psychologists rely primarily on self-report outcome measures to validate scores derived from psychological tests (Bornstein, 2001), they frequently discuss validity evidence based on self-reports as if this evidence reflected actual behavior (Bornstein, 2003). As numerous researchers have noted, most studies using traditional validity assessment procedures fail to examine differential test score validities in different contexts and settings (Mischel, 1984; Mischel, Shoda, & Mendoza-Denton, 2002), in different groups of respondents (Young, 2001), and as a function of the stated or implied purpose of the test (Steele & Aronson, 1995).

Psychology's reliance on correlational methods to quantify validity would be problematic in the best of circumstances, but is especially so given the nature of our discipline: As Meehl (1978) observed, in the social sciences, everything correlates with everything, at least to some degree. Meehl's wry observation echoes an earlier conclusion by Guilford (1946) that summarized in stark terms the fundamental problem with the traditional correlational-observational conceptualization of validity: When validity is equated with magnitude of predictor–criterion association, a test score is, by definition, valid for anything with which it correlates.

Psychology can do better. By using experimental manipulations to alter respondents' psychological processes during testing and assessing the impact of these manipulations on test scores, strong conclusions can be drawn regarding whether or not a test score is actually measuring what it is thought to measure. When these data are combined with traditional predictor–criterion association results, validity assessment will become more rigorous, the utility of psychology's measurement procedures will be enhanced, and test bias can be minimized.

A PF Model of Validity

In the PF model, validity is conceptualized as the degree to which respondents can be shown to engage in a predictable set of psychological processes during testing; once these processes are identified, experimental manipulations are introduced to alter these processes and determine whether the manipulations affect test scores in meaningful ways. The PF framework reverses the usual procedure for dealing with extraneous variables that alter psychological test scores: Rather than regarding them as problematic, the PF model conceptualizes variables that are seen as confounds in traditional validity assessment (e.g., self-presentation effects) as opportunities for manipulation, exploration, and focused analysis—windows on underlying processes that would otherwise remain hidden.

Instrument-Based Processes

To a substantial degree, the psychological activities in which people engage when responding to psychological tests are determined by the nature of the instruments themselves—by the types of questions asked and the tasks and activities required of the respondent. Table 3 uses a process-based framework to classify the array of assessment tools used by psychologists today, grouping these instruments into six categories based on the mental activities and behaviors involved in responding to these tests (see also Bornstein, 2007, for a detailed discussion of this issue).[5]

As Table 3 shows, s*elf-attribution tests* (which are usually described as objective or self-report tests; e.g., the NEO Personality Inventory; Costa & McCrae, 1985) typically take the form of questionnaires wherein people are asked to acknowledge whether or not each of a series of descriptive statements is true of them, or rate the degree to which these statements describe them accurately. *Stimulus attribution tests* require people to interpret ambiguous stimuli, and here the fundamental task is to attribute meaning to a stimulus that can be interpreted in multiple ways; this attribution process occurs in much the same way as the attributions that each of us make dozens of times each day as we navigate the ambiguities of the social world (e.g., when we attempt to interpret our friend's failure to greet us as we pass on the street; see Kawada, Oettingen, Gollwitzer, & Bargh, 2004).

Performance-based tests include the Bender (1938) Visual-Motor Gestalt Test; the Implicit Association Test (Nosek, Greenwald, & Banaji, 2005); occupational screening tools that require behavior samples as part of the assessment; and various intelligence, mental state, and neuropsychological measures. Within the PF framework, *performance-based tests* are distinguished from stimulus-attribution tests because different processes are involved: Whereas performance-based tests require the respondent to perform structured behavioral tasks (e.g., copy figures from cards, assemble jigsaw puzzles), with performance evaluated according to predefined scoring criteria, respondents' scores on stimulus-attribution tests like the Rorschach Inkblot Method (Rorschach, 1921) and Thematic Apperception Test (Murray, 1943) are derived from open-ended descriptions and elaborations of test stimuli. In the PF framework, *constructive tests* are also distinguished from stimulus-attribution tests, because constructive tests require respondents to create—literally to "construct"—novel products (e.g., drawings, written descriptions) with minimal guidance from the examiner and no test stimulus physically present (e.g., Machover's, 1949, Draw-a-Person [DAP] test). In contrast to stimulus-attribution tests, which require respondents to describe stimuli whose essential properties were determined a priori, in constructive tests the "stimulus" exists only in the mind of the respondent (e.g., a self-schema or parental image).

Continuing through Table 3, *observational measures* (as are often used to quantify behavior in hospitals, classrooms, shopping malls, and other settings; e.g., Baltes, 1996; Sproull, 1981), may be distinguished from *informant-report tests* (wherein data are derived from knowledgeable informants' descriptions or ratings; e.g., Achenbach, Howell, Quay, & Conners, 1991). Though in both cases, judgments are made by an individual other than the person being evaluated, different processes are involved in generating these judgments, with observational measures based on direct observation and immediate recording of behavior, and informant-report tests based on informants' retrospective, memory-derived conclusions regarding characteristics of the target person (see Meyer et al., 2001, for a discussion of self- vs. informant-derived psychological test data).

Context-Based Influences

In contrast to instrument-based processes, which are inherent in the measure itself, context-based influences are situational factors that alter test responses by influencing respondents' motivations and goals, or by modifying aspects of respondents' cognitive or emotional processes during testing. Context-based influences not only reflect external variables that affect test performance (historically conceptualized as confounds to be minimized) but also represent potential manipulations that—when used to alter instrument-based processes in theoretically meaningful ways—provide unique information regarding the mental operations that occur during testing. Context-based influences may be divided into four categories.

Assessment setting. Assessment setting effects are shaped not only by the physical milieu in which testing occurs (e.g., corporate office, psychiatric hospital, research laboratory) but also by respondents' perceptions of, and beliefs regarding, this milieu (see Butcher, 2002, and Rosenthal, 2003, for examples). Thus, a student who had a pleasant experience in an earlier learning disability (LD) assessment is likely to approach testing more openly—less defensively—than one whose past LD assessment experiences were negative. A person voluntarily seeking admission to a psychiatric unit will respond to self-attribution test items in an intake packet quite differently than a person who has been brought to the unit involuntarily. Someone completing psychological tests as part of their induction into the military is likely to approach testing very differently if they have been drafted than if they volunteered for service.

Instructional set. Studies have demonstrated that the way an instrument is labeled and described influences the psychological processes that occur during testing. For example, Steele and Aronson (1995) found that African American—but not Caucasian—college students perform more poorly on Scholastic Aptitude Test (SAT) items when these items are identified as indices of intelligence than when the same items are identified as indices of problem-solving ability; presumably the increased anxiety experienced by African American students who are concerned that their performance might confirm a preexisting racial stereotype diminishes attentional capacity and temporarily impairs certain cognitive skills. Using a very different paradigm and set of outcome measures, Bornstein, Rossner, Hill, and Stepanian (1994) found that college students' self-attributed interpersonal dependency scores increased when testing was preceded by a positive description of dependency-related traits and behaviors, but decreased when testing was preceded by a negative description of dependency. These same students' stimulus-attribution-based dependency scores (i.e.,

[5] Portions of this section are adapted from Bornstein (2007, pp. 203–204).

Table 3
A Process-Based Framework for Classifying Psychological Tests

Test category	Key characteristics	Representative tests
Self-attribution	Test scores reflect the degree to which the person attributes various traits, feelings, thoughts, motives,behaviors, attitudes, or experiences to him- or herself.	NEO Personality Inventory Strong Vocational Interest Blank Beck Depression Inventory
Stimulus-attribution	Person attributes meaning to an ambiguous stimulus, with attributions determined in part by stimulus characteristics and in part by the person's cognitive style, motives, emotions, and need states.	Rorschach Inkblot Method Thematic Apperception Test
Performance based	Test scores are derived from person's unrehearsed performance on one or more structured tasks designed to tap on-line behavior and responding.	Wechsler Adult Intelligence Scale Bender Visual-Motor Gestalt Test
Constructive	Generation of test responses requiresperson to create or construct a novel image or written description within parameters defined by the tester.	Draw-a-Person Test Qualitative and Structural Dimensions of Object Relations
Observational	Test scores are derived from observers'ratings of person's behavior exhibited in vivo, or in a controlled setting.	Spot Sampling Behavior Trace Analysis
Informant report	Test scores are based on knowledgeable informants' ratings or judgments of a person's characteristic patterns of behavior and responding.	SWAP-200 Informant-Report version of the NEO Personality Inventory

Note. Adapted from Table 1 in "Toward a Process-Based Framework for Classifying Personality Tests: Comment on Meyer and Kurtz (2006)" by R. F. Bornstein, 2007, *Journal of Personality Assessment, 89,* pp. 202–207. Copyright 2007 by Taylor & Francis. Reprinted with permission.

scores on Masling, Rabie, & Blondheim's, 1967, Rorschach Oral Dependency [ROD] scale) were unaffected by instructional set.[6]

Affect state. A respondent's emotional state (e.g., elated, depressed, anxious) affects test responses in at least two ways. First emotional reactions—especially strong emotional reactions—take up cognitive capacity, making it more difficult for the person to focus attention on the task at hand or divide their attention between competing tasks (Arnell, Killman, & Fijavz, 2007). In this way, emotional reactions alter performance on measures of intelligence, aptitude, achievement, neuropsychological functioning, and mental state. Second, moods and other affect states have biasing effects, priming affect-consistent nodes in associative networks and thereby increasing the likelihood that certain associates and not others will enter working memory (Hänze & Meyer, 1998; Robinson & Clore, 2002). Studies have also shown that people are more likely to retrieve mood-congruent than mood-incongruent episodic memories, though these effects are stronger when free-recall procedures are used than when highly structured (e.g., questionnaire) measures are used (McFarland & Buehler, 1998; Zemack-Ruger, Bettman, & Fitzsimons, 2007). Thus, mood-priming effects are particularly salient when stimulus-attribution tests and constructive tests are administered.

Examiner effects. As Masling (1966, 2002) and others have shown, examiner characteristics and behaviors alter psychological test responses in predictable ways (see Butcher, 2002, for reviews of studies in this area). For example, testers who interact with respondents in a distant or an authoritarian manner elicit self-attribution and stimulus-attribution test responses that are more guarded and defensive than those elicited by testers who treat respondents more warmly during the evaluation. When examiners create rapport with respondents prior to administering intelligence test items, respondents tend to produce higher intelligence scores than are obtained when testing is not preceded by rapport building. Similar findings emerge in performance-based occupational screens. Garb (1998) provided an extensive review of the literature documenting the impact of clinician expectancy effects on the outcome of psychological assessments. Although some of these biasing effects stem from clinicians' misperceptions of respondent behavior based on characteristics of the person being evaluated (e.g., gender, age, physical attractiveness), these effects also stem from the manner in which the examiner interacts with the examinee prior to and during testing, which may alter the examinees' cognitive processes, emotional states, and motives (Allen, Montgomery, Tubman, Frazer, & Escovar, 2003).

Implementing the PF Model

Table 4 summarizes in broad terms the four steps involved in test score validation using the PF model. As Table 4 shows, the first step in process-focused test score validation involves specifying the underlying processes that should occur as individuals respond to test stimuli (e.g., retrospective memory search, associative priming) and identifying context variables (e.g., affect state, instructional set) that potentially alter these processes. Next, process–outcome links are operationalized and tested empirically (Step 2), and the results of these assessments are evaluated (Step 3). Finally, process-focused test score validity data are contextualized by enumerating limiting conditions (e.g., flaws in experimental design) that might have influenced the results and evaluating the generalizability and ecological validity of PF data by

[6] Although the RIM has been the topic of considerable controversy in recent years, much of this debate has centered on the utility of Exner's (1991) comprehensive system. Even vocal critics of the RIM acknowledge the psychometric soundness of the ROD scale and the strong validity evidence in support of the measure. As Hunsley and Bailey (1999) noted, "One excellent example of a scale that does have scientific support . . . is the Rorschach Oral Dependency scale. The history of research on this scale may serve as a useful guide for future attempts to validate [other] Rorschach scales" (p. 271).

Table 4
A Process-Focused Model of Validity

1) Deconstruct assessment instrument(s)
a) Specify underlying psychological processes
b) Identify context variables that alter these processes
2) Operationalize and evaluate process–outcome links
a) Turn process-altering variables into manipulations
b) Delineate hypothesized outcomes
c) Experimental design
3) Interpret outcome
a) Process-based validity results
b) Limiting conditions
4) Evaluate generalizability/ecological validity
a) Population
b) Context and setting

assessing the degree to which similar patterns are obtained in different populations and settings. This latter task entails conducting replications of the initial investigation in different contexts, using new participant samples. Thus, Steps 1–3 will occur whenever a process-focused validity study is conducted; Step 4 represents a long-term goal that requires additional studies.

Research examining the process-focused validity of scores derived from self-attribution and stimulus attribution measures of interpersonal dependency illustrates one way in which the PF model may be implemented. As Bornstein (2002) noted, the Interpersonal Dependency Inventory (IDI; Hirschfeld et al., 1977) and ROD scale (Masling et al., 1967) are both widely used, and both yield well-validated (from an outcome perspective) scores that have been shown to predict a broad array of dependency-related behaviors (e.g., suggestibility, help seeking, compliance, interpersonal yielding) in laboratory and field settings (see Bornstein, 1999, for a meta-analysis of behaviorally referenced validity evidence for these two measures). Although scores derived from the IDI and ROD scale show good concurrent and predictive validity, IDI and ROD scale scores correlate modestly with each other—typically in the range of .2 to .3—raising questions regarding the degree to which the two measures are tapping similar constructs (see McClelland, Koestner, & Weinberger, 1989, for parallel findings regarding the intercorrelation of self-attribution and stimulus attribution need for achievement scores).

Using the logic of the PF model, Bornstein et al. (1994) and Bornstein, Bowers, and Bonner (1996a) assessed the process validity of IDI and ROD scale scores in a series of experiments wherein manipulations were used to alter one set of processes but not the other, and assess the differential impact of these manipulations on self- versus stimulus attribution dependency scores. In the first investigation, Bornstein et al. (1994) deliberately altered participants' self-presentation goals by introducing an instructional manipulation immediately prior to testing. Bornstein et al. administered the IDI and ROD scale to a mixed-sex sample of college students under three different conditions. One third of the participants completed the tests in a *negative set* condition; prior to completing the IDI and ROD scale, these participants were told that both were measures of interpersonal dependency and that the study was part of a program of research examining the negative aspects of dependent personality traits (following which several negative consequences of dependency were described). One third of the participants completed the two measures in a *positive set* condition; these participants were told that the study was part of a program of research examining the positive, adaptive aspects of dependency (following which several positive features of dependency were described). The remaining participants completed the measures under standard conditions, wherein no mention is made of the purpose of either scale or the fact they assess dependency. Bornstein et al. (1994) found that relative to the control condition, participants' IDI scores increased significantly in the positive set condition and decreased significantly in the negative set condition; ROD scores were unaffected by instructional set.

In a follow-up investigation, Bornstein et al. (1996a) used a retest design to examine the impact of induced mood state on IDI and ROD scores, having college students complete the two measures under standard conditions, then calling participants back for a second testing 6 weeks later and asking them to write essays regarding traumatic events, joyful events, or neutral events to induce a corresponding mood immediately prior to testing. On the basis of previous findings regarding the impact of mood on the priming of nodes in associative networks (e.g., Rholes, Riskind, & Lane, 1987), Bornstein et al. hypothesized that induction of a negative mood state would produce a significant increase in dependent imagery (e.g., increases in associations related to passivity, helplessness, frailty, and vulnerability), leading to increases in ROD scale scores. Because the impact of mood on response to questionnaire items is comparatively modest (Hirschfeld, Klerman, Clayton, & Keller, 1983), Bornstein et al. hypothesized that IDI scores would not increase significantly in the negative mood condition. The expected patterns were obtained: Induction of a negative mood led to a significant increase in ROD—but not in IDI—scores (see Bornstein, 2002, for descriptions of other PF studies involving measures of interpersonal dependency).

Thus, the PF model proved useful in illuminating the processes that underlie self-attribution and stimulus-attribution dependency scores, and in helping explain the modest intercorrelations between scores on two widely used measures of the same construct (see also McClelland et al., 1989, for a discussion of this issue). Similar logic can be used in other domains as well. For example, one might examine the psychological processes involved in responding to self-attribution narcissism test items by manipulating respondents' self-focus (e.g., inducing self-focus vs. external/field focus using a mirror manipulation; see George & Stopa, 2008) prior to testing. By introducing a cognitive load as participants complete a brief neurological test or dementia screen, evidence regarding the process-focused validity of these measures can be examined. To ascertain whether state anxiety can indeed be inferred from DAP test data (Briccetti, 1994), an anxiety-inducing manipulation (vs. no manipulation) can be implemented. Finally, given the psychological processes that occur during observational ratings and informant reports, one might expect that providing false feedback regarding an individual prior to obtaining observer and/or informant judgments regarding that person would alter these judgments in predictable ways. Because observational ratings occur in the here-and-now whereas informant reports are retrospective (and therefore more susceptible to retrieval-based memory distortion), one would hypothesize that false feedback should alter informant reports more strongly than observational ratings.

Implications of the PF Model: Research, Psychometrics, Practice, and Social Policy

Table 5 contrasts the traditional and PF models in five areas: evidence, research strategy, validity coefficient generalizability, test development goals, and challenges. In addition to highlighting operational differences between the two perspectives, Table 5 illustrates how the PF model shifts psychologists' understanding of the generalizability of validity data (from concordance of validity coefficients across groups to documentation of similar underlying processes in different groups), and the strategies involved in test development (from finding optimal criterion measures and maximizing test score–criterion relationships to finding optimal manipulations and maximizing the impact of these manipulations on underlying process).

Although the PF model yields unique information that the traditional outcome-focused approach cannot provide, neither method alone yields a truly comprehensive picture of test score validity. When both approaches are used, psychologists can derive two separate validity coefficients, both of which may be represented as standard effect sizes (e.g., *r* or *d*): an *outcome effect size* (the traditional estimate of predictor–criterion association), and a *process effect size* (a numerical index of the degree to which a theoretically relevant manipulation altered test score in line with a priori predictions). Moreover, just as one may conceptualize the outcome effect size as a single predictor–criterion correlation or as the sum total (or average) of an array of interrelated predictor–criterion correlations (Rosenthal, 1991), one may conceptualize the process effect size with respect to a single experimental manipulation, or an array of converging manipulations that would all be expected to have similar effects on a given test score.

Note that when the two frameworks are integrated in this way, a given measure can potentially fall into one of four categories:

1. *Adequate outcome and process validity*. This is the best possible result, reflecting a situation wherein a test score predicts what one hopes it does, and the psychological processes in which respondents engage while completing the measure are in line with expectations.

2. *Adequate outcome but not process validity*. In this situation, the test score appears to be assessing what one hopes, but it is not clear why, because respondents' reactions during testing differ from what was expected.

3. *Adequate process but not outcome validity*. Here, the measure seems to be tapping the expected underlying psychological processes, but test scores do not relate to external, theoretically related indices as anticipated.

4. *Inadequate outcome and process validity*. Neither process nor outcome are as expected, and it might be time to move on to a new test.

As these four scenarios illustrate, a key advantage of combining outcome and process validity data is that these data not only point to potential limitations in a measure but also suggest specific interventions to correct these limitations. Scenario 2 suggests devoting greater attention to process than outcome issues, and determining whether the problematic process results reflect difficulties in the test itself or the manipulation used to evaluate it. Scenario 3 suggests devoting greater attention to outcome than to process issues, and considering whether the disappointing predictor–criterion relationships stem from flaws in the test or in the outcome measures used to validate scores derived from it.

Thus, the PF model represents both an affirmation of and challenge to the 1999 *Standards'* conceptualization of validity as a unitary concept, with validity broadly defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association et al., 1999, p. 9). In support of this view, the PF model suggests that a complete picture of test score validity can only be obtained by integrating divergent sources of validity data obtained via different methods and procedures. In contrast to the unitary concept view, however, the PF model argues that distinctions among certain types of validity evidence (in this case, process validity and outcome validity) remain useful and should be retained. Process and outcome validity evidence for a given test score should be considered both separately and in combination so that convergences and divergences among different forms of validity data can be scrutinized. The notion of validity as a truly unitary concept—though admirable—is premature.

Research and Psychometric Implications

Although implementing the PF model involves shifting the focus of validity research from correlation to experimentation, process-focused validity studies need not be limited to true experiments wherein underlying processes are manipulated directly, but may also involve quasiexperiments wherein preexisting groups are selected on the basis of presumed process differences. For example, following up on their initial investigations, Bornstein, Bowers, and Bonner (1996b) found significant positive correlations between IDI scores and Bem's (1974) Sex Role Inventory (BSRI) femininity scores, and significant negative correlations between IDI scores and BSRI masculinity scores. Stimulus attribution dependency scores were unrelated to gender role orientation in

Table 5
Contrasting the Traditional and Process-Focused Models

Domain	Traditional model	Process-focused model
Key evidence	Degree to which test score correlates with theoretically related variable	Degree to which test score is altered via manipulation of theoretically related process
Research method	Assessment of predictor–criterion correlation	Assessment of impact of experimental manipulation
Validity coefficient generalizability	Concordance of validity coefficients across groups	Documentation of similar processes across groups
Test development goals	Maximize test score–outcome correlation	Demonstrate impact of theoretically related process
Test development challenges	Finding optimal criterion measure(s)	Finding optimal manipulation(s)

participants of either gender. These patterns suggest that gender differences in self-reported dependency are due, at least in part, to women's and men's efforts to present themselves in gender-consistent ways on psychological tests. Not surprisingly, a meta-analytic synthesis of gender differences in self-attribution and stimulus attribution dependency scores found that women scored significantly higher than men on every self-attribution dependency scale, but not on any stimulus attribution dependency measure (Bornstein, 1995).

The PF model has implications for test development, in that potential test stimuli must not only be constructed to maximize the association between test score and theoretically related variables, but also with respect to underlying process issues. In other words, test items should be constructed to engage those psychological processes (e.g., retrospection, spontaneous association) that reflect the construct being assessed and the method being used to assess it. Thus, in addition to performing preliminary factor- and cluster-analytic studies when refining psychological test items, and evaluating the degree to which potential test items are associated with external indices, psychometricians should assess the impact of relevant process manipulations on responses to each test item.

One might reasonably argue that requiring psychometricians to conduct experimental process-focused validity studies prior to the publication of psychological tests puts an undue burden on test publishers, slowing the test development process and delaying the introduction of new measures that could potentially benefit patients, clinicians, and psychologists in various applied settings (e.g., organizational, forensic, etc.). There is no ideal solution to this dilemma, and the best approach may be one that seeks a middle ground: Just as a substantial (but not necessarily comprehensive) body of psychometric data should be obtained before a new psychological test is used in vivo, a substantial (but not necessarily definitive) body of process-focused experimental evidence should be collected prior to publication of a new test. Moreover, just as researchers continue to collect psychometric data postpublication so the strengths and limitations of a test can be better understood and the measure revised and improved, researchers should continue to collect process-focused data postpublication so the underlying processes engaged by the measure are brought into sharper focus. These data can also be used to refine and improve the test.

As Bornstein et al.'s (1996a, 1994) findings illustrated, inherent in the PF model is a new conceptualization of test score divergence: When two measures of a construct engage different underlying processes, one can deliberately dissociate these processes, using manipulations that alter one set of processes but not the other. Thus, in addition to shifting the emphasis from outcome to process, and from correlation to experimentation, the PF model calls our attention to the importance of meaningful test score discontinuity (see also Meyer, 1996, 1997, and Meyer et al., 2001, for discussions of this issue). Just as convergent validity evidence must be accompanied by discriminant validity evidence to yield a complete picture of test score validity using the traditional outcome-based approach, manipulations that cause two measures of a given construct to converge more strongly must be complemented by manipulations that cause scores on these measures to diverge more sharply when the PF model is used. Note that manipulations that cause two test scores to converge should also increase the convergence of these two test scores with a common theoretically related external criterion, whereas manipulations that cause two test scores to diverge should lead to greater divergence in the magnitude of test score–criterion links. This is another means through which process-focused and outcome-based validity data may be integrated.

Practice and Social Policy Implications

Principle 9.05 of the American Psychological Association's (2002) *Ethical Principles and Code of Conduct* states that "Psychologists who develop tests and other assessment techniques use appropriate psychometric procedures and current scientific or professional knowledge for test design, standardization, validation, reduction or elimination of bias, and recommendations for use" (p. 14). The PF model's framework for conceptualizing test score divergence has implications for understanding the sources of group differences in performance; ultimately the PF model may enhance psychologists' ability to develop measures that generalize more effectively across gender, age, race, and ethnicity, thereby reducing test bias and test score misuse.

Although public controversy regarding test bias has tended to emphasize group differences in outcome (e.g., ethnic and racial differences in SAT scores), psychometricians have increasingly focused on differential predictor–criterion relationships as a key index of bias (e.g., situations wherein scores on a personnel selection screen predict occupational success more effectively in members of one group than another). The PF model provides a framework for evaluating the degree to which group differences in predictive validity may be rooted in underlying process: When a test score predicts an outcome more effectively in one group than in another, this differential outcome validity is likely to be rooted, at least in part, in intergroup process differences, and these can be detected by introducing manipulations designed to alter the processes in question. Thus, differential predictive validities of SAT scores in African American and Caucasian students should increase when manipulations design to increase stereotype threat are used, and decrease when threat-reducing manipulations are used. Gender differences in self-attributed dependency should increase when test instructions are written to focus respondents' attention on gender role issues, and decrease when instructions that deemphasize gender are used (see Major & O'Brien, 2005, for a discussion of contextual cues that moderate self-schema- and self-presentation-related psychological processes in various groups).

Two practice and policy implications follow, one having to do with addressing concerns regarding test bias prior to publication, the other with remedying flaws in existing tests. With respect to the former, psychologists developing measures that have historically tended to yield problematic group differences should deliberately evaluate the degree to which similar underlying processes are engaged in different groups, using standard PF manipulations (e.g., changes in test labels, induction of a negative mood or state anxiety) during the early stages of item development. When process differences are identified, these can be addressed before the test is used in vivo (e.g., by altering item content, revising test instructions, or evaluating the impact of varying item formats on performance).

With respect to the latter issue, the PF model suggests an alternative definition of test bias: empirically demonstrable differences in the psychological processes engaged by different groups

of respondents. With this in mind, educators, policymakers, and mental health professionals who seek to document test bias (or the absence of bias) can use a process-focused framework alongside the traditional outcome-based approach. Once process-based sources of bias are identified in research settings, strategies for reducing these sources of bias in vivo may be implemented. In forensic contexts, demonstrable group differences in process—when coupled with differences in predictor–criterion relationships— represent compelling evidence that an assessment procedure does not yield comparable outcomes in different groups.[7]

Conclusion: Toward an Integrated, Integrative Perspective on Test Score Validity

The goals of psychology have evolved during the past several decades, and so must the goals of validity research. Historically the relationship between experimentation and test score validity has been largely unidirectional, as researchers sought instruments with well-validated scores to enhance the rigor of their experiments. The PF model turns this unidirectional relationship into a bidirectional one: Just as one cannot conduct a rigorous experiment without valid test scores, one cannot validate test scores rigorously unless one uses experimental procedures as part of the overall validation strategy.

Unlike traditional outcome-based validity assessment, the PF model explicitly links psychological testing to other areas of psychology (e.g., cognitive, social, developmental). Many of the manipulations used in PF studies to date have drawn upon ideas and findings from psychology's various subfields, including research on memory, mood, self-presentation, implicit motivation, gender role socialization, and other areas. In this respect, the PF model not only enhances psychologists' understanding of test score validity, but may also help connect psychology's disparate subfields, contributing to the unification of a discipline that has fractionated considerably in recent years.

[7] The PF model has pedagogical implications as well, teaching students the value of experimental methods and the ways in which experimental data enrich correlational results. Because the PF framework links assessment to other areas of psychology (e.g., cognitive, social), it helps deepen students' perception of psychological science as a unified discipline. Moreover, the phenomenological emphasis of the PF framework—increased attention to the mental processes and subjective experience of the respondent—enables students to grasp the complexities of psychological assessment in ways that the traditional approach cannot.

References

Achenbach, T. M., Howell, C. T., Quay, H. C., & Conners, C. K. (1991). National survey of problems and competencies among four- to sixteen-year olds: Parents' reports for normative and clinical samples. *Monographs of the Society for Research in Child Development, 56*(3, Serial No. 225).

Allen, A., Montgomery, M., Tubman, J., Frazier, L., & Escovar, L. (2003). The effects of assessment feedback on rapport-building and self-enhancement process. *Journal of Mental Health Counseling, 25,* 165–182.

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

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct.* Washington, DC: Author.

Arnell, K. M., Killman, K. V., & Fijavz, D. (2007). Blinded by emotion: Target misses follow attention capture by arousing distractors in RSVP. *Emotion, 7,* 465–477. doi:10.1037/1528-3542.7.3.465

Baltes, M. M. (1996). *The many faces of dependency in old age.* Cambridge, England: Cambridge University Press.

Bem, S. L. (1974). The measurement of psychological androgeny. *Journal of Consulting and Clinical Psychology, 42,* 155–162. doi:10.1037/h0036215

Bender, L. (1938). *A visual-motor gestalt test and its clinical use.* New York, NY: American Orthopsychiatric Association.

Bollen, K. A., & Davis, W. R. (2009). Causal indicator models: Identification, estimation, and testing. *Structural Equation Modeling, 16,* 498–522. doi:10.1080/10705510903008253

Bornstein, R. F. (1995). Sex differences in objective and projective dependency tests: A meta-analytic review. *Assessment, 2,* 319–331. doi:10.1177/1073191195002004003

Bornstein, R. F. (1999). Criterion validity of objective and projective dependency tests: A meta-analytic assessment of behavioral prediction. *Psychological Assessment, 11,* 48–57. doi:10.1037/1040-3590.11.1.48

Bornstein, R. F. (2001). Has psychology become the science of questionnaires? A survey of research outcome measures at the close of the 20th century. *The General Psychologist, 36,* 36–40.

Bornstein, R. F. (2002). A process dissociation approach to objective-projective test score interrelationships. *Journal of Personality Assessment, 78,* 47–68. doi:10.1207/S15327752JPA7801_04

Bornstein, R. F. (2003). Behaviorally referenced experimentation and symptom validation: A paradigm for 21st century personality disorder research. *Journal of Personality Disorder, 17,* 1–18.

Bornstein, R. F. (2007). Toward a process-based framework for classifying personality tests: Comment on Meyer and Kurtz (2006). *Journal of Personality Assessment, 89,* 202–207.

Bornstein, R. F., Bowers, K. S., & Bonner, S. (1996a). Effects of induced mood states on objective and projective dependency scores. *Journal of Personality Assessment, 67,* 324–340. doi:10.1207/s15327752jpa6702_8

Bornstein, R. F., Bowers, K. S., & Bonner, S. (1996b). Relationships of objective and projective dependency scores to sex role orientation in college students. *Journal of Personality Assessment, 66,* 555–568. doi:10.1207/s15327752jpa6603_6

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment, 63,* 363–386. doi:10.1207/s15327752jpa6302_14

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071. doi:10.1037/0033-295X.111.4.1061

Briccetti, K. A. (1994). Emotional indicators of deaf children on the Draw-a-Person test. *American Annals of the Deaf, 139,* 500–505.

Butcher, J. N. (Ed.). (2002). *Clinical personality assessment: Practical approaches* (2nd ed.). New York, NY: Oxford University Press.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105. doi:10.1037/h0046016

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand-McNally.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68,* 397–412. doi:10.1177/0013164407310130

Costa, P. T., & McCrae, R. R. (1985). *NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D.

(2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences, 33,* 137–150. doi:10.1017/S0140525X09991567

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. doi:10.1037/h0040957

Embretson (Whitely), S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197. doi:10.1037/0033-2909.93.1.179

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York, NY: Plenum Press.

Embretson, S. (1998). A cognitive design system approach for generating valid tests: Approaches to abstract reasoning. *Psychological Methods, 3,* 300–396.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38,* 343–368.

Exner, J. E. (1991). *The Rorschach: A comprehensive system, Volume 2* (2nd ed.). New York, NY: Wiley.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment.* Washington, DC: American Psychological Association. doi:10.1037/10299-000

George, G., & Stopa, L. (2008). Private and public self-awareness in social anxiety. *Journal of Behavior Therapy and Experimental Psychiatry, 39,* 57–72. doi:10.1016/j.jbtep.2006.09.004

Goldman, B. A., & Mitchel, D. F. (2003). *Directory of unpublished experimental mental measures* (Vol. 8). Washington, DC: American Psychological Association.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36,* 181–192.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427–438.

Hagemann, D., & Meyerhoff, D. (2008). A simplified estimation of latent state-trait parameters. *Structural Equation Modeling, 15,* 627–650. doi:10.1080/10705510802339049

Hänze, M., & Meyer, H. A. (1998). Mood influences on automatic and controlled semantic priming. *American Journal of Psychology, 111,* 265–278. doi:10.2307/1423489

Hershberger, S. L. (2003). The growth of structural equation modeling: 1994–2001. *Structural Equation Modeling, 10,* 35–46. doi:10.1207/S15328007SEM1001_2

Hirschfeld, R. M. A., Klerman, G. L., Clayton, P. J., & Keller, M. B. (1983). Personality and depression: Empirical findings. *Archives of General Psychiatry, 40,* 993–998.

Hirschfeld, R. M. A., Klerman, G. L., Gough, H. G., Barrett, J., Korchin, S. J., & Chodoff, P. (1977). A measure of interpersonal dependency. *Journal of Personality Assessment, 41,* 610–618. doi:10.1207/s15327752jpa4106_6

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64,* 802–812. doi:10.1177/0013164404264120

Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment, 11,* 266–277. doi:10.1037/1040-3590.11.3.266

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical tests of the rule assessment methodology by latent class analysis. *Developmental Review, 17,* 321–357.

Jansen, B. R. J., & Van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81,* 383–416. doi:10.1006/jecp.2002.2664

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58,* 736–753. doi:10.1177/0013164498058005002

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kawada, C. L. K., Oettingen, G., Gollwitzer, P. M., & Bargh, J. A. (2004). The projection of implicit and explicit goals. *Journal of Personality and Social Psychology, 86,* 545–559. doi:10.1037/0022-3514.86.4.545

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55,* 95–107.

Machover, K. (1949). *Personality projection in the drawing of the human figure.* Springfield, IL: Charles C Thomas. doi:10.1037/11147-000

Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. *Annual Review of Psychology, 56,* 393–421. doi:10.1146/annurev.psych.56.091103.070137

Masling, J. M. (1966). Role-related behavior of the subject and psychologist and its effect upon psychological data. In D. Levine (Ed.), *Nebraska symposium on motivation* (pp. 67–104). Lincoln: University of Nebraska Press.

Masling, J. (2002). Speak, memory, or goodbye, Columbus. *Journal of Personality Assessment, 78,* 4–30. doi:10.1207/S15327752JPA7801_02

Masling, J., Rabie, L., & Blondheim, S. H. (1967). Obesity, level of aspiration, and Rorschach and TAT measures of oral dependence. *Journal of Consulting Psychology, 31,* 233–239. doi:10.1037/h0020999

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96,* 690–702. doi:10.1037/0033-295X.96.4.690

McFarland, C., & Buehler, R. (1998). The impact of negative affect on autobiographical memory: The role of self-focused attention to moods. *Journal of Personality and Social Psychology, 75,* 1424–1440. doi:10.1037/0022-3514.75.6.1424

McIntire, S. A., & Miller, L. A. (2000). *Foundations of psychological testing.* Boston, MA: McGraw-Hill.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834. doi:10.1037/0022-006X.46.4.806

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13–23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as inquiry into score meaning. *American Psychologist, 50,* 741–749. doi:10.1037/0003-066X.50.9.741

Meyer, G. J. (1996). The Rorchach and MMPI: Toward a more scientific understanding of cross-method assessment. *Journal of Personality Assessment, 67,* 558–578. doi:10.1207/s15327752jpa6703_11

Meyer, G. J. (1997). On the integration of personality assessment methods: The Rorschach and MMPI. *Journal of Personality Assessment, 68,* 297–330. doi:10.1207/s15327752jpa6802_5

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Read, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165. doi:10.1037/0003-066X.56.2.128

Mischel, W. (1984). Convergences and divergences in the search for

consistency. *American Psychologist, 39,* 351–364. doi:10.1037/0003-066X.39.4.351

Mischel, W., Shoda, Y., & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science, 11,* 50–54. doi:10.1111/1467-8721.00166

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36,* 463–469. doi:10.3102/0013189X07311660

Murray, H. A. (1943). *Thematic Appreciation Test manual*. Cambridge, MA: Harvard University Press.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: Method variables and construct validity. *Personality and Social Psychology Bulletin, 31,* 166–180. doi:10.1177/0146167204271418

Rholes, W. S., Riskind, J. H., & Lane, J. W. (1987). Emotional states and memory biases: Effects of cognitive priming and mood. *Journal of Personality and Social Psychology, 52,* 91–99. doi:10.1037/0022-3514.52.1.91

Robinson, M. D., & Clore, G. L. (2002). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology, 83,* 198–215. doi:10.1037/0022-3514.83.1.198

Rorschach, H. (1921). *Psychodiagnostik*. Bern, Switzerland: Bircher.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Thousand Oaks, CA: Sage.

Rosenthal, R. (2003). Covert communication in laboratories, classrooms, and the truly real world. *Current Directions in Psychological Science, 12,* 151–154. doi:10.1111/1467-8721.t01–1-01250

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Mahwah, NJ: Erlbaum.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Slaney, K. L., & Maraun, M. D. (2008). A proposed framework for conducting data-based test analysis. *Psychological Methods, 13,* 376–390. doi:10.1037/a0014269

Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Sproull, L. S. (1981). Managing education programs: A micro-behavioral analysis. *Human Organization, 40,* 113–122.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811. doi:10.1037/0022-3514.69.5.797

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5,* 1–25.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1,* 31–65. doi:10.1146/annurev.clinpsy.1.102803.144239

Ullman, J. B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment, 87,* 35–50. doi:10.1207/s15327752jpa8701_03

Young, J. W. (2001). *Differential validity, differential prediction, and college admissions testing: A comprehensive review and analysis.* New York, NY: The College Board.

Zemack-Rugar, Y., Bettman, J. R., & Fitzsimons, G. J. (2007). Effects of nonconsciously priming emotional concepts on behavior. *Journal of Personality and Social Psychology, 93,* 927–939. doi:10.1037/0022-3514.93.6.927

Received July 1, 2010
Revision received November 12, 2010
Accepted November 15, 2010 ■