The Meaning of Personality Test Scores

Robert Hogan and Robert A. Nicholson
University of Tulsa

ABSTRACT: The process of reexamining the methodological and metatheoretical assumptions of personality psychology over the past two decades has been useful for both critics and practitioners of personality research. Although the field has progressed substantially, some critics continue to raise 1960s-vintage complaints, and some researchers perpetuate earlier abuses. We believe that a single issue—construct validity—underlies the perceived and actual shortcomings of current assessment-based personality research. Unfortunately, many psychologists seem unaware of the extensive literature on construct validity. This article reviews five major contributions to our understanding of construct validity and discusses their importance for evaluating new personality measures. This review is intended as a guide for practitioners as well as an answer to questions raised by critics. Because the problem of construct validity is generic to our discipline, these issues are significant not only for personality researchers but also for psychologists in other domains.

Walter Mischel’s 1968 book stimulated 20 years of careful examination of the methodological and metatheoretical assumptions of personality psychology. This reappraisal has been useful for both critics and practitioners of measurement-based personality research. Within the personality research establishment, things seem to have returned to normal; people are comfortable once again with the notions that (a) personality assessment is an appropriate methodology in many areas of personality, clinical, and industrial psychology, and (b) there is stability to personality descriptors over time and occasions. Nonetheless, plus ça change, plus c’est la même chose: On one hand, 1960s-vintage criticisms continue to come up (i.e., test scores are contaminated by social desirability; validity coefficients are modest; there is no stable core to personality). On the other hand, some personality researchers continue to invite reproof by perpetuating abuses about which critics legitimately complain.

Landy (1986) documented the confusion that still surrounds the testing enterprise despite 20 years of soul searching. We believe that a single issue underlies the perceived and actual shortcomings of current assessment-based personality research. When critics complain and when personality researchers fail, construct validity is usually at the heart of the matter. A review of the notion of construct validity may serve as a guide for practitioners as well as an answer to many of the questions that seem to bother critics of measurement-based personality research (cf. Kagan, 1988). Drawing on Landy’s essay, in this article we argue that (a) all validity is construct validity; (b) the process of test validation is hypothesis testing; and (c) measurement-based research is formally identical with any other type of legitimate scientific inquiry.

In its most general case, measurement-based research in personality takes the following form: (a) The researcher hypothesizes that individual differences in position D are associated with variations in performance of a certain type P (e.g., variations in attitudes toward authority are associated with variations in conformity to social norms); (b) the researcher chooses a measure of disposition D and an index of performance P; (c) the researcher selects a sample, assesses individuals with the measure of D, and records for each person in the sample an index of his or her performance, P. Finally, the researcher examines the extent to which scores on the measure of disposition D covary with the index of P.

If the scores on the measure of D reliably covary with the index of P, then the researcher has established an empirical relationship, but one that requires further investigation and amplification: Will the finding hold up for different samples? Will other “variables” moderate the relationship (e.g., is it stronger for men than for women)? However, what we believe to be the most crucial question is not asked often enough: What does the covariation mean? The question is important because we have not compared D with P directly, and the meaning of the relationship is therefore obscure. Traditional logical empiricism, the scientific perspective of choice for generations of psychologists, maintained that prediction and explanation have the same logical form (cf. Hempel, 1965). Many psychologists seem to be turning away from this older viewpoint and arguing that the act of theory building goes beyond the (still crucial) process of establishing reliable covariations (cf. Fiske & Shweder, 1986). We agree. If scores on the measure of D covary with an index of performance P, then that is interesting information in itself, but it is only the first step in what should be a lengthy process of inquiry, data gathering, statistical analysis, and hypothesis formation. As Landy (1986) noted, the distinction between the meaning and the predictive value of test scores remains crucial.

For some time, . . . validity was considered a correlation between a predictor and a criterion . . . Such a positivist view was (and remains) only minimally helpful in developing . . . a basic understanding of what was being measured . . . [Consequently] the emphasis on the meaning of test scores is as important today as it was 30 years ago. (p. 1183)
On the other hand, what can we conclude if scores on the measure of D are unrelated to the index of performance P? There are four possibilities: (a) The initial hypothesis was wrong; (b) the hypothesis was right, but the research sample was inappropriate; (c) the measure of D does not reflect the disposition in question (e.g., attitudes toward authority); or (d) the index of P does not reflect performance in the domain of interest (e.g., conformity to social norms). If the researcher uses measures of D and P that are well-supported by other empirical findings, then conclusions (c) and (d) can be to some degree ruled out, and knowledge has still been advanced. If, as sometimes happens, the investigator uses a measure of D that is not "well-validated" and there is no covariation with the index of P, then one can conclude very little. Unfortunately, the literature is replete with examples of researchers testing substantive hypotheses with homemade and unvalidated scales; when it is later discovered that the scales did not measure what they purported to measure, the entire line of research is called into question (cf. Briggs & Cheek, 1988; Hansson, Hogan, Johnson, & Schroeder, 1983).

How are we to avoid these kinds of embarrassments? We can avoid them by taking the concept of validity seriously, and that, in turn, requires understanding construct validity. As Landy (1986) and others have argued, all questions about test validity ultimately concern construct validity; the various types of validity identified by the American Psychological Association (APA) Committee on Test Standards are special cases of construct validity; the various types of validity identified by the American Psychological Association (APA) Committee on Test Standards are special cases of construct validity; the various types of validity identified by the American Psychological Association (APA) Committee on Test Standards are special cases of construct validity. Cronbach and Meehl (1955) introduced the notion of construct validity to modern psychology in an effort to solve the problem of criterion selection in test validation. Their classic article provided the foundation for subsequent elaborations of the kinds of research needed to establish construct validity. Briefly, they noted that for many interesting psychological concepts: (a) no single criterion can serve as primary validation for a particular test or scale; and (b) one can always ask whether a single criterion is in fact a valid indicator of the construct in question. This second point leads to an infinite regress as one provides a criterion to validate a scale score, then a second order criterion to validate the first criterion, then a third order criterion to validate the second, and so on. Because no single criterion is sufficient to establish the validity of a measure, Cronbach and Meehl (1955) recommended examining the entire network of external correlates of a test. Careful reflection on this network of correlates provides information useful for understanding a test's construct validity; in Kagan's (this issue, pp. 614-620) terms, the pattern of referential meanings serves to illuminate the sense of a concept as reflected in a standard psychological measure. Cronbach and Meehl (1955) argued that the process of test validation is in principle open-ended; it consists of developing and evaluating a theory about what scores on a test actually mean. In this sense, the process of construct validation is formally identical with the investigation and elaboration of the nomological network in which the construct is embedded.

Loevinger (1957) proposed three kinds of considerations for evaluating the degree to which test A is a valid indicator of disposition D. First she recommended examining the "substantive components" of a test. This term refers to the item contents and universes included and excluded from a particular scale or test. Is the item content appropriate to disposition D? Does the item content adequately survey the domain of D? Are all of the items equally relevant or does the test include items that are more appropriate to another domain? Second, Loevinger suggested considering the "structural components" of the measure. One's theory about disposition D will allow one to make certain inferences about the correlations among the items on test A (the measure of D), and about how these item intercorrelations might resemble patterns of covariation among indexes of performance P. 

Correspondence concerning this article should be addressed to Robert Hogan, Department of Psychology, University of Tulsa, 600 South College Ave., Tulsa, OK 74104.

August 1988 • American Psychologist
Finally, Loevinger proposed evaluating the "external components" of the measure. What are the correlations between the items and total scores on test A and indexes of performance P? These recommendations may seem obvious to many readers, but the dust-bowl empiricist tradition of scale construction paid little or no attention to item content, adequacy of domain sampling, or item covariations.

Campbell and Fiske (1959) suggested combining convergent and discriminant criteria to answer construct validity questions. They recommended that researchers obtain scores for more than one disposition using more than one method of assessment. For example, subjects might complete personality inventories, Q sorts, and a role-playing task, and they also might be described by objective raters and by friends using adjective checklists. The researcher can then construct a correlation matrix using scores on all of these procedures. Assume that the procedures described above provided four different measures of four dispositions, A through D. To support the notion that disposition D exists, the four measures of that disposition should be more highly correlated with one another than with any of the measures of dispositions A, B, and C. Convergent validity (e.g., substantial correlations) among the measures of disposition D supports the inference that D exists, but does not prove it. The Campbell–Fiske article extended the analysis of construct validity in two ways. On the one hand, it provides a statistical mechanism for distinguishing trait-variance from method variance, which is believed to artificially inflate correlations between measures of a disposition using the same methodology. On the other hand, the Campbell–Fiske model allows us to determine what a particular test measures and what it does not measure. For example, a valid measure of attitudes toward authority ought not to be a measure of extraversion, and the knowledge that it is not is useful information pertinent to the construct validity of the original measure.

As Landy (1986) pointed out, when applied psychologists use tests, they are interested in the inferences they can make about a person based on his or her test scores. This is the point of Gough's (1965) contribution to the construct validity literature. Gough proposed three kinds of procedures through which the meaning of test scores can be appraised. At the first level, one examines the external correlates of scale scores: With what is the test correlated and with what is it not correlated? A second level of analysis broadens the scope of the validity investigation to include personological implications of test performance. At this level, one evaluates the interpersonal significance of a scale score by determining how persons with high and low scores on the scale are perceived by others. At the third level, one looks for serendipitous and unexpected covariations with nontest behavior.

Gough's example will work well here; consider the Socialization scale of the California Psychological Inventory (CPI; Gough, 1987). The scale was designed to identify individual differences in the willingness to comply with the ordinary norms and conventions of society. At the first level of validity evaluation, then, the scale should distinguish between delinquents and nondelinquents, and in fact, it does this with considerable power and reliability. At the second level of validity evaluation, persons with high and low scores should create characteristic but differing impressions in those who know them. Again, this turns out to be the case: Persons with high scores on the Socialization scale are described as mature, responsible, and calm, whereas persons with low scores are described as hostile, impulsive, and immature. At the third level of evaluation one looks for unanticipated but interpretable relationships between test scores and nontest performance. In the case of the CPI Socialization scale, it turns out that among otherwise normal adults, low scores on socialization are associated with rated and demonstrated creativity. Gough's analysis extends the notion of validity considerably beyond the range of the initial correlates of a scale score. In effect, Gough (1965) redefined validity to include the number and range of valid inferences a user can make about a particular client on the basis of a test score; he also suggested a set of systematic procedures that an investigator can follow in order to evaluate the validity of a measure.

Buss and Craik (1983a, p. 1083 ff.) took a different approach to the construct validity problem by providing an act frequency analysis of criterion development. Their approach may offer the best answer so far to the question, Does test A assess disposition D? Briefly, the procedure consists of asking people to nominate acts (specific intentional behaviors) that reflect the operation of a disposition D. The nominated acts are compiled, edited, and then judged for prototypicality—certain acts will be more prototypical of disposition D than others. By this procedure, an open-ended and extensive domain of prototypical acts for a dispositional construct can be generated and identified. The ability to formulate a consensual list of prototypical acts is prima facie evidence that disposition D exists. Observed frequencies of prototypical acts in everyday conduct then provide criterion data against which scale scores from test A can be compared. This comparison yields an estimate of the degree to which test A is a measure of disposition D. Thus, the existence of multiple acts organized around a prototype allows inferences of an underlying attribute; correlations between ratings based on these acts and test A allow the inference that A is a valid measure of the dispositional construct (see also Buss & Craik, 1983b).

Taken together, the discussions by Cronbach and Meehl, Loevinger, Campbell and Fiske, Gough, and Buss and Craik tell us how to evaluate the construct validity of a test. Not all measures will be able to pass all of the validity hurdles just described. For example, Graduate Record Examination (GRE) scores have few correlations with nontest behaviors other than academic test performance. GRE scores also permit few inferences about test takers, and they have few correlates with other assessment procedures. In a slightly different vein, the Jenkins Activity Survey (JAS) seems modestly predictive of some kinds of cardiovascular problems (Jenkins, Zyzanski, &
Rosenman, 1979), but our understanding of (a) what scores on the JAS mean, and (b) the Type A construct ostensibly tapped by the instrument is woefully inadequate (Hansson, Hogan, Johnson, & Schroeder, 1983).

The primary problem with measurement-based personality research, it seems to us, is the tendency of naive researchers to introduce measures and do some preliminary validation research (e.g., examine the prediction of a criterion), but to fail to ask what scores on the test mean or what inferences can be made on the basis of the test scores. Such researchers often seem to think more carefully about a test's name than the construct that a test is designed to measure, and therefore they fail to pursue more sophisticated validation procedures. It is somewhat paradoxical that, after all the recent serious discussion of personality measurement, many psychologists seem unaware of the literature on construct validity adequate (Hansson, Hogan, Johnson, & Schroeder, 1983).

Rosenman, 1979), but our understanding of (a) what scores on the JAS mean, and (b) the Type A construct ostensibly tapped by the instrument is woefully inadequate (Hansson, Hogan, Johnson, & Schroeder, 1983).

The primary problem with measurement-based personality research, it seems to us, is the tendency of naive researchers to introduce measures and do some preliminary validation research (e.g., examine the prediction of a criterion), but to fail to ask what scores on the test mean or what inferences can be made on the basis of the test scores. Such researchers often seem to think more carefully about a test's name than the construct that a test is designed to measure, and therefore they fail to pursue more sophisticated validation procedures. It is somewhat paradoxical that, after all the recent serious discussion of personality measurement, many psychologists seem unaware of the literature on construct validity adequate (Hansson, Hogan, Johnson, & Schroeder, 1983).

We have suggested that the literature on construct validity can answer many of the questions that seem to bother critics of measurement-based personality research (e.g., Kagan, this issue, pp. 614–620). There are, however, four other common criticisms that might be useful to discuss briefly here. These concern social desirability, correlations with physiological measures, the interpretation of “self-report” data, and the view that relationships between personality descriptions only reflect the semantic links between trait words. Although these criticisms of personality measurement have been largely rebutted, the message somehow seems not to get across to some writers.

Consider first the “problem” of social desirability. Edwards (1953) noted that the rated social desirability of Minnesota Multiphasic Personality Inventory (MMPI) items correlated about .80 with their endorsement frequencies in a normal population. He concluded that people respond to the desirability of items rather than their content, and this set off a controversy that brought personality assessment research to a near standstill for almost 10 years. Subsequent research shows that Edwards’s data are correct but that his conclusion is less persuasive. For example, Block (1965) demonstrated, using the MMPI item pool, that the factor structure of the MMPI remains essentially unchanged whether or not one controls for the social desirability of the items. This, in turn, suggests that people primarily respond to the content of items, regardless of their rated desirability. In a more pragmatic vein, Dicken (1963), McCrae and Costa (1983), and McCrae (1986) reasoned that if social desirability response set contaminates item endorsements, then controlling for this “response bias” should enhance the validity of scale scores. These investigators provided converging and replicated data showing quite clearly that when social desirability is controlled, validity coefficients are instead reduced. From a historical perspective, it now appears that the social desirability response set hypothesis is most usefully seen as an early theory of item responses rather than an indictment of personality assessment.

Other critics of measurement-based personality research complain about the lack of stable covariations between personality measures and measures of corresponding physiological processes; they argue, for example, that scores on anxiety scales should correlate with physiologically based measures of anxiety. Because reliable covariations of this type are relatively rare, they question the validity of “self-report” scales. In principle, we believe that neurological and neurochemical processes parallel personality processes in some very complex way, to argue otherwise would be to take an obscurantist view of personality research. Nonetheless, it seems strange to assume that people can and will accurately report on their physiological activity under all circumstances. Averill (1982) and Sarbin (in press) argued persuasively that emotions are best regarded as social rather than physiological phenomena. Specifically, there are normative expectations regarding emotional performances. For example, one feels sad at funerals, reverent in church, and jubilant in victory. It is a short step from the normative to the strategic performance of emotion; people can use emotional performances to promote interpersonal agendas. Consequently, from a dramaturgical or role-theoretical perspective, one would never expect to find simple parallels between test scores and physiological measures of variables with the same names.

Also implicit in this criticism is the assumption that if various indicators of a hypothetical construct are unrelated, then the indicators may not measure the same construct, or to use Kagan’s (this issue, pp. 614–620) terms, a change in reference may change the sense of the construct. However, that assumption is not necessarily valid. An adequate theory will specify the conditions under which the indicators of a particular construct can be expected to covary, but sometimes the theory will predict conditions under which covariations should not be observed. For example, a role-theoretical perspective on emotion would predict the finding that physiological and self-report indexes of anxiety correspond for novice but not for expert parachutists (cf. Fenz & Epstein, 1967).

The role of theory for interpreting covariations among the indicators of a construct is not limited to the relationship between physiological and self-report measures. Bollen (1984) showed, for example, that for certain theoretical models, the indicators of a latent entity may have negative, positive, or zero correlations with one another. For example, marital satisfaction and length of marriage are positive indicators of marital stability, even though marital satisfaction is negatively correlated with length of marriage. The key to this paradox appears to be whether the measured variables are considered to be cause-indicators or effect-indicators of the construct in question. Consequently, it is not necessary that we always find positive correlations between different measures of a construct, although one’s theory should specify the conditions under which the various outcomes will occur.

Our third point concerns the term self-report, as in “self-report measures.” Naive test users often take item endorsements at face value and treat them as if they were a second-best way to observe behavior. If they are assumed
to be behavior samples that have some degree of correspondence with actual behavior, then these item responses can also be used as criterion data to validate other scale scores and measures. Critics of measurement-based personality research object to this simplistic validation strategy, and for good reason. However, if item responses are not self-reports, then what are they? Assume that you have been asked to assist your city in hiring a new chief of police. Assume that you have agreed to do this and that you are now interviewing the candidates and asking them about their philosophy of management, attitudes toward minorities, and the role of women in law enforcement. How will you interpret their answers to your questions? Are they factual summaries of past behavior called up from memory? Or are they efforts to project a particular image that you, the interlocutor, will find acceptable, efforts that are guided by an idealized self-image that the candidates believe and hope you will too?

Paul Meehl (1945) argued in a different context that item endorsements are interesting bits of verbal behavior whose nontest meanings remain to be determined. We would modify this statement in light of our police chief interview example. Responses to items on questionnaires are not self-reports; they are self-presentations that are formally identical to those that characterize answers to questions during an employment interview. Even so, the meanings of these self-presentations are not transparent. We know that when they are appropriately aggregated, they often allow us to predict various kinds of social behavior, such as conformity, extraversion, shyness, impulsivity, and management potential. These item endorsements are not self-reports, however, and it is misleading to conceptualize them in this way. Despite the fact that these self-presentations can be put to predictive purposes, their meaning is a matter for subsequent analysis and investigation.

Another frequent criticism of personality measurement rests on the view that observers are unable, in principle, to rate the personality characteristics of actors and that ratings reflect the manner in which personality terms are organized in the raters' semantic memory rather than characteristics of the actor's behavior. This view, known as the "semantic distortion hypothesis," was initially proposed by D'Andrade (1965), cited approvingly by Mischel (1968), and subsequently amplified by Berman and Kenny (1976), Bourne (1977), and especially Shweder (1982; Shweder & D'Andrade, 1980). The argument has serious implications for personality psychology; Shweder (1975), for example, asked, "How relevant is an individual differences theory of personality" (p. 453), and replied that he had "shown" that "an individual difference theory of personality [is] no more than statements about how respondents classify things as alike in meaning" (p. 452). The data and the logic on which the semantic distortion hypothesis depend have been seriously questioned (Block, Weiss, & Thorne, 1979; Romer & Revelle, 1984; Rowe, 1982), but Weiss and Mendelsohn (1986) provided the most damaging critique. They showed, in a parallel set of analyses, one contaminated and one uncontaminated by semantic distortion effects, that target persons are described in the same way in both analyses. Most observers regard the semantic distortion hypothesis as having been largely disproved (see also DeSoto, Hamilton, & Taylor, 1985).

A final word is in order. Our discussion and examples have concerned personality assessment and research. Some readers might take this to mean that the problem of construct validity is unique to measurement-based personality research. It is not. Consider, for example, studies of experimental psychopathology. Under controlled laboratory conditions, researchers deprive animals of preferred physical and social stimulation—food, water, sleep, contact with other animals—and then manipulate characteristics of the environment—predictability and controllability—by altering discriminative stimuli, consequences, and specific contingencies. The behaviors that follow such manipulations are frequently characterized by increased activity, defecation, and repetitive performance of previously reinforced responses. Are experimental psychopathologists primarily interested in the relation between food deprivation, environmental contingencies, and behavior, and nothing more? Or, is their real concern with the relationships among "need frustration," "competence frustration," and the subsequent "driven-ness" and "rigidity" that characterize neurotic behavior?

To the extent that such animal models are useful analogues of human psychopathology, experimentalists must be concerned with construct validity. In any case, as Garner, Hake, and Eriksen (1956) pointed out over three decades ago, the meaning of scores on dependent variables in experimental research is as ambiguous as it is in assessment research.

It pleases the adepts of certain religious orders (Cistercian monks, born-again Christians, Moslem fundamentalists) to believe that they have discovered a unique path to truth. Nonbelievers tend to be skeptical about the claim that these people have discovered a privileged access to knowledge. Within academic psychology, researchers also have their preferred methods. Some prefer survey data, some prefer meta-analysis, some prefer assessment data, and some prefer data obtained from controlled experimentation. None of these methods yields canonical results. All depend on indirect measures and a statistical methodology that yields probabilistic results at best, and all are subject to the challenges posed by construct validity.

To speak of any construct loosely, that is, outside the context of the laws, theoretical entities, and observables that give the construct meaning, is to fall into what Meehl (1978) called "fuzzy verbalism." On the other hand, to attend only to an operational indicator of a construct is to engage in "fake operationalism." In this sense, the process of construct validation is akin to sailing between Scylla and Charybdis. We must continually work back and forth between the indicators and the construct without being swallowed up by either. Fortunately, the contributions to construct validation described in this article can guide our course. The problem of construct
validity is generic to our discipline, and it is only partly solved by attending to the referential meanings of our constructs.

REFERENCES


