ABSTRACT: There have been continuing arguments with respect to the unitarian versus trinitarian view of the validation process. An emerging position is that the threefold distinction among validation models presented by Cronbach and Meehl is not as useful as it once was. This article joins in the unitarian argument and suggests that the validation process be considered nothing more and nothing less than traditional hypothesis testing. The issue of the role of constructs in psychological measurement is considered, as is the general status of construct-validation strategies in employment testing. In addition, the unique knowledges, skills, and abilities that psychologists bring to the discussion of validity and inference are discussed.

During the past 15 years, there have been many occasions to regret the simple taxonomy of validity strategies suggested by the American Psychological Association in their "Technical Recommendations for Psychological Tests and Diagnostic Techniques" (1954). Cronbach and Meehl were members of the APA committee that developed those recommendations. They further elaborated on the notion of construct validity in a subsequent publication in 1955 (Cronbach & Meehl, 1955). The "Technical Recommendations" proposed a four-category taxonomy of validity models. The taxonomy consisted of predictive, concurrent, content, and construct validity. The predictive and concurrent models are usually combined into a category known as "criterion-related validation strategies." Thus, it is common to hear reference made to three "types" of validity. This view of validity has been aptly labeled the trinitarian view (Guion, 1980).

At the time that the "Technical Recommendations" were published, the taxonomy was quite valuable. It provided order where before there was chaos. It added the preliminary structure necessary for progress. For some time, this taxonomy proved a useful tool for validity discussions. The taxonomy represented a point of departure. Prior to that point, validity was considered a correlation between a predictor and a criterion. It was a literalist view directed toward the technology of prediction. Such a positivist view was (and remains) only minimally helpful in developing theories of behavior or a basic understanding of what was being measured by the tests in question. The introduction of the "Technical Recommendations" and the concepts of content-oriented and construct validation represented an attempt to correct the past excesses of dust-bowl empiricism and focus attention on understanding rather than on the simpler and less meaningful search for predictability.

The value of that reorientation has been substantially blunted in the past 15 years as a result of a reification of the validity taxonomy that was introduced in 1954. This result is unfortunate because the emphasis on the meaning of test scores rather than the predictive value of those scores is as important today as it was 30 years ago. As a result of the efforts of practitioners (lawyers, judges, expert witnesses, and personnel administrators) to make the taxonomy into something it was never intended to be, the point of departure suggested by Cronbach and Meehl (1955) has become a point of arrival. The three types of validity that were originally proposed as a method for broadening the appreciation of the inferential limits of test scores have become an orthodoxy.

Because forensic activity is usually deductive in nature, there has been the tendency in Title VII litigation to try to identify the validity model most appropriate for a given situation. Once the validity model has been identified, the case revolves around whether the requirements of the model have been met. In principle, this is no different from the way in which any court case is approached: Determine which principle covers the issue under scrutiny, and then determine if that particular principle has been violated. This approach can be quite effective if the principles are clear and distinct from one another and if all the parameters of the principles are accepted by all the parties involved. In psychology, an example of the value of this approach has been the development of the DSM-III for the diagnosis of psychological disorders (American Psychiatric Association, 1980). Although it is not without its detractors, the DSM-III is a useful way of determining from a set of observable symptoms what might be responsible for those symptoms and how those symptoms might be best treated. The parameters of the diagnostic model are clearly defined and understood (the five "axes" of the DSM-III), and there is a clear listing of the symptoms that warrant one diagnosis as opposed to another.

Many would like to use the Uniform Guidelines on Employee Selection Procedures (1978) in a similar manner. Unfortunately, there is substantial disagreement about the meaning and the independence of the principles involved in determining the legality of certain classes of employment decisions. The principles are, for the most part, validity "models." However, if these models are accepted as "principles" in a literal sense, the result is a
form of forensic and scientific myopia in which all parties wittingly or unwittingly agree to limit their discussion to the boundaries suggested by the validity models. Viewed from another perspective, this willingness to accept a small and fixed set of validity models is nothing more than a procrustean worldview in which some aspects of the conceptual issue are lopped off and thrown away and other natural conceptual limits are stretched well beyond their elasticity index. Tenopyr (1984) recently bemoaned this point:

Certainly that division has rendered more mischief than practically anything I know in the area of test theory. Despite the fact that the framers of the tripartite division were cautious and spoke of things like “aspects” of validity, as was done in the Standards for Educational and Psychological Tests in the 1974 edition, the fact that a distinction was made among content, construct and criterion-related validity has led these three concepts to be reified and split much further apart than their framers intended. (pp. 3–4)

The latest revision of the Standards for Educational and Psychological Testing (1985) suggests that there is still some ambiguity regarding the nature of validation. On the one hand, the Standards suggest that the three categories are simply convenient labels. Thus, after introducing the three traditional labels of content-related evidence, criterion-related evidence, and construct-related evidence, the Standards (1985) go on to state the following:

These categories are convenient, as are other more refined categorizations (e.g., the division of the criterion-related category into predictive and concurrent evidence of validity) but the use of the category labels does not imply that there are distinct types of validity or that a specific validation strategy is best for each specific inference or test use. (p. 9)

On the other hand, in the following section, Standards 1.6, 1.10, 1.11, 1.16, 1.19, and 1.20 clearly invoke the traditional models and the “convenient label” concept is not substantially advanced by these references. In a later section on employment testing (pp. 59–62), the extent of the recidivism is apparent. Standards 10.1, 10.4, 10.5, 10.6, 10.7, and 10.8 clearly rest on the content–criterion–construct distinction. The addition of the suffix “related” to each of these concepts does not address the problem. This suffix is intended to emphasize the fact that tests do not “possess” validity. It does nothing to dispel the notion of the three “acceptable” models of validation.

Validity Models and the Uniform Guidelines

The consequences of this reification are substantial. Either directly or indirectly, many Title VII decisions are influenced by a judge’s decision with respect to whether the correct model has been chosen. The jeopardy inherent in this labeling process is eloquently expressed in Guardians Association v. Civil Service Commission of the City of New York (1980). One aspect of the litigation involved the plaintiffs’ assertion that a particular test measured “constructs” and, as a result, the only appropriate validation strategy would be a construct-validation approach. The validity strategy chosen by the defendants was a content-oriented approach. In writing the decision for the Second Circuit Court of Appeals, Judge Newman clearly identified the pivotal issue: “This content–construct distinction has a significance beyond just selecting the proper technique for validating the exam; it frequently determines who wins the lawsuit” (p. 92).

As a result of this tendency to label validity approaches as correct or incorrect in a given situation, Title VII cases often take on the appearance of a primitive form of stamp collecting. There are only three spaces to be filled—the content space, the construct space, and the criterion-related space. The test in question is the metaphorical stamp. If it is a test of constructs, then it is pasted in the construct space, and the litigants set out to determine if all of the requirements for construct-validation efforts have been met. If it is a test of knowledges, skills, or abilities currently possessed by an applicant, then either the content or criterion-related space is filled and the litigants consider a different checklist of requirements. These list checkers take on the appearance of the modern-day equivalent of the biblical Pharisees, checking scripture to determine if law or tradition has been violated.

Lines of Defense

Three lines of reasoning have been advanced as defenses against the stamp collectors. The first might be labeled the separate and equal defense. Until the 1978 revision of the Uniform Guidelines, criterion-related validity was considered to be the ideal model, and content-oriented strategies were occasionally acceptable but seldom respected. With the 1978 revision of the Uniform Guidelines, equal status was given to content-oriented strategies. In addition, construct-oriented strategies were considered acceptable but defined in such a way as to emphasize, once again, the more equal status of criterion-related validation efforts. It was suggested that any construct-related effort must include a criterion-related study as well. Note that this defense does not deny the propriety of three models or categories; it simply argues that one should not predominate.

The second line of defense, like the first, does not really deny the propriety of three models of validity. Instead, it places emphasis on the relationships among the models. This line of reasoning might be labeled the continuum defense. An example of this defense can be found in the propositions of Tenopyr (1977, 1984), who argued that construct-oriented strategies are really subordinate and that criterion-oriented or content-oriented strategies are subordinate. In other words, content- and criterion-oriented strategies are part of (and may, in fact, help to

---

I gratefully acknowledge the assistance of Amy Schwartz in identifying literature related to the role of constructs in psychological measurement. In addition, I am grateful to Ted Abel, Wayne Caska, Robert Guion, Rick Jacobs, and Steve Motowidlo for comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Frank J. Landy, Department of Psychology, Pennsylvania State University, University Park, PA 16802.

November 1986 • American Psychologist
define) the broader construct-oriented strategy. (Cronbach, 1985; Messick, 1980). A popular variation on this theme has been that one type of strategy (e.g., content-oriented) is not always easily distinguished from another type of strategy (e.g., construct-oriented) and that the strategies actually lie on an abstract–concrete continuum rather than existing as discrete strategies or models. The recent Standards for Educational and Psychological Testing (1985) seem to accept, at least implicitly, this line of reasoning by suggesting that "rigorous distinctions between the categories (of validation strategies) are not possible" (p. 9).

Recently, a third line of defense has appeared. Lawshe (1985) has noted that types of validity do not exist. Instead, one should consider types of analysis. In that sense, it can be inferred that Lawshe is suggesting an emphasis on designs, analyses, and subsequent inferences as a system rather than considering each of these three components in isolation. However, even Lawshe inadvertently supported the trinitarian view by suggesting that there are three types of analysis that can be used to support a validation study. These types of analysis produce criterion-related, content, and construct evidence. As a result, it is not clear if his suggestion represents a genuine contribution to the discussion.

Lawshe’s suggestion that validation be viewed as a system is excellent advice. None of these components (i.e., designs, analyses, or inferences), taken in isolation, is protection against the stamp collectors. Similarly, each of the three defenses admits to the viability of the three-category taxonomy. Fortunately, there is an emergent trend toward a consideration of validation as a multidimensional process, as Lawshe implied. It is abundantly clear that this process is tied to inference.

Types of Inference

The major point that seems to emerge from the discussions of models of validity is that certain (validation) processes permit certain inferences. What is not always apparent in these discussions is that there are different sorts of inferences that can be drawn from a given set of test scores. Consider the following inferences:

- (a) X is a part of Y.
- (b) X is an approximation of Y.
- (c) X is a sign of Y.

Inference a is a simple instance of extension. If X is a work sample and Y is the job, then the inference is that a person who does well on a representative sample of the job will do well on the job as a whole. In the trinitarian model, this would be a minimally serviceable definition of content-oriented test validation. Inference b says that X is an example of Y. This might represent a rough working definition of construct validity. Inference c says that Y can be predicted from X, an acceptable representation of a criterion-oriented validation strategy in which X is the test score and Y is the performance score. The point of this description of different inferences is that the labels content, construct, and criterion-related are not completely useless, nor are they interchangeable. They had their value in 1954, and they have their value in 1986. However, their value is not as types of validity. Instead, their value is in pointing out that there is more than one type of inference that can be made from a test score.

Multiple Inferences in Construct Validation

One of the problems that has emerged in the discussion of constructs is directly related to the issue of inferences as just outlined. Consider Proposition b again. In the original formulation of the notions of construct validation, a duality of purpose was implied. One could focus on identifying an underlying attribute that was responsible for observed individual differences, or one could attempt to determine if a particular test was a good measure of an underlying attribute with some established integrity. This duality of purpose was well represented in the attempt by Campbell and Fiske (1959) to operationalize the construct-validation process. Through their multitrait multimethod framework, they suggested that the integrity of the construct could be affirmed by looking at the extent to which different measures of the same attribute converged. Similarly, the meaning of the construct was further illuminated by showing that method variance was not a viable hypothesis for explaining the intercorrelations among different measures of the same attribute. However, just as researchers used the multitrait multimethod matrix to solidify the interpretation of a particular construct, they also used it to choose the best method of measuring that construct. Best was commonly defined as the method with the strongest pattern of convergent validities and discriminant validities and with the highest reliability.

To return to the discussion of inferences outlined earlier, Inference b can now be further subdivided into Inferences b1 and b2, along the lines implied by the multitrait multimethod strategy:

- (b1) Y is an attribute that can account for differences in X.
- (b2) X requires the same attribute for success as Y.

Loevinger noted this distinction in 1957 and described b1 as the validity of the construct and b2 as the validity of the test for the construct. Whitely (1983) recently expanded on that point and used the term construct representation to describe b1 and the term nomothetic span to describe b2. Her point was that sometimes one is interested in whether a construct is useful at all (b1), and sometimes one is interested in whether the test being used is a measure of the construct of interest (b2). Whitely made an important observation about paradigm shifts in psychology that may provide a clue as to why problems are emerging in the attempt to apply the original labels of the Technical Recommendations (American Psychological Association, American Educational Research Association, & National Council of Measurement Used in...
Education, 1954) to the current theory and practice in applied measurement. In the early 1950s, the prevailing paradigm was one of positivism. The concern was for accounting for observable variables by considering antecedent and consequent conditions. The paradigm shift that has occurred in virtually all of psychology has been a heightened concern for the unobservable conditions. This shift has been particularly evident in the area of cognitive behavior. Whitely pointed out that this shift has redefined what are relevant data and, as a result, has re-oriented the types of questions that are asked in designing validity studies. In short, she suggested that the discipline is a good deal more dependent on constructs now than it may have been 30 years ago.

The thread that runs through the original propositions of the “Technical Recommendations,” the elaboration of Cronbach and Meehl (1955), the extension of Loevinger (1957), the operationalization of Campbell and Fiske (1959), and the recent considerations of construct validity (e.g., Whitely, 1983) is that in all discussions of validity, the concern is with the issue of the extent to which inferences can be justified.

The theme of inference introduced earlier in the context of general measurement and theory construction makes the point that researchers are not really interested in the properties of tests. Instead, they are interested in the attributes of people who take those tests. Thus, validation processes are not so much directed toward the integrity of tests as they are directed toward the inferences that can be made about the attributes of the people who have produced those test scores (Cronbach, 1971; Dunnette & Borman, 1979; Guion, 1980). Parenthetically, however, it should be noted that the use of the term test validity is often a form of psychological metonymy in which a sign replaces the thing signified. In fact, in many forensic settings, an expert witness is well advised to establish quickly the distinction between the validity of inferences and the validity of tests and then accept the shorthand nomenclature of judges and lawyers regarding the “validity of tests.” To do otherwise would be to substitute a phrase such as “a set of observations” for the phrase “a set of test scores.”

Validation as Hypothesis Testing

Even though the emphasis of recent discussions of models of validity has been directed toward inferences rather than tests (Cronbach, 1985; Lawshe, 1985; Standards, 1985; Tenopyr, 1984), bolder steps must be taken to break away from the trinitarian doctrine of validity (Guion, 1980). Perhaps a new direction can be developed by welding the notion of a validation process to the notion of inferences from test scores.

In Lawshe’s (1985) argument in favor of types of validity analyses rather than types of validity, he defined validity analysis as “a procedure, a process or a strategy whereby we collect or generate data to determine or defend the extent, degree or strength of the inference or inferences that can be made from a set of test scores” (p. 237). With only minor modifications, this definition can be easily transformed into a more general definition of hypothesis testing. The only critical modification would be to substitute a phrase such as “a set of observations” for the phrase “a set of test scores.” In viewing the definition in this light, several things become obvious. The first is that the number of validity analyses available is limited only by the creativity and experience of the analyst. There are certainly more than three. In addition, it should be obvious that the three categories of validity analysis suggested by Cronbach and Meehl (1955) are convenient caricatures of a much richer set of procedures available to the analyst. Granted, these caricatures may be modal, but they no more describe the range of strategies available than a mean describes the range of scores in a distribution.

As illustrative examples of variations of validity design that have been discussed recently, consider the “aspects” of test validity suggested by Messick (1980). His design departed from the three traditional designs by showing that within each of these designs, many subdesigns exist. Each subdesign permits (or supports) a slightly different inference that one might attempt to make from a test score. For example, in the generic category of construct validity, Messick included the following subcategories: convergent validity, discriminant validity, trait validity, nomological validity, factorial validity, substantive validity, structural validity, external validity, population validity, ecological validity, temporal validity, and task validity. In a similar vein, Guion and Cranny (1982) identified at least five variations on the basic predictive-validity design. They also suggested at least three designs for a concurrent-validity study. Messick (1980) and Guion and Cranny (1982) emphasized the design characteristics of a validity study. They considered the types of subjects, the nature of the data-gathering procedures, the variables involved, and the inferences that could potentially be drawn from the results.

The validity analyst is carrying out traditional hypothesis testing. At least by implication, the hypothesis being considered is of the following form: People who do well on test X will do well on activity Y, or \( Y = f(X) \). In null form, the hypothesis would assert that there is no relationship between test performance and activity performance.

If that hypothesis were the core of a master’s thesis or doctoral dissertation, I suspect that there would be minimal discussion of the right type of validity. There would be discussion of whether the study should be a lab or field study, which subjects would be most appropriate, what conceptual and operational definitions would be adopted for the variables of interest, what controls (statistical or experimental) should be imposed, what analytic strategies would be used to test the hypothesis, and finally, what inferences might be drawn from the results (particularly
with respect to generalization), and what caveats would be necessary in light of possible threats to accurate inference.

The Industrial/Organizational Psychologist as Scientist

In the language of job analysis, "validation behavior" should be carried out by individuals with the requisite knowledge and ability. Industrial and organizational psychologists should be well suited to validation activities. The requisite knowledge is the body of information related to principles of human behavior. This knowledge makes a psychologist better suited than a biologist, an astronomer, or a statistician to address the challenge of predicting human behavior. The requisite ability is related to the present capacity to design a study that will support or refute the generic test performance-job performance hypothesis stated earlier. Viewed in this light, the labels attached to validity designs become not only less limiting but also less threatening. No one complains when designs are labeled as longitudinal or cross-sectional, field study or laboratory experiment, or nested or completely crossed. When reasonable people share a concept and a language, there is no problem. The concepts are not asked to do more than that for which they were intended. Unfortunately, in the context of Title VII discussions, similar labels (content, criterion-related, construct) are used as cudgels rather than as tools dedicated to conceptual efficiency. Thus, although there should be no reason to fear such labels as content, construct, or criterion-related as limiting conditions, experience indicates the opposite.

At the root of the problem, there seems to be a confusion between the aim of employment testing and methods to be used for validating inferences from employment test scores. It is becoming more widely recognized that these two issues are distinct. Simply because the goal is to draw inferences about future job performance from present test scores, it does not necessarily follow that criterion-related validity strategies are the only means for documenting the soundness of those inferences. The point is well made in several recent statements. Consider two of them:

Prediction is often thought of as closely associated with criterion-related validation. In employment situations, the use of scores from a procedure developed on the basis of content also has a predictive basis. That is, one measures performance in a domain of job activities that will be performed later. (American Psychological Association, Society for Industrial and Organizational Psychology, 1980, p. 5)

It has been suggested that a test validated by criterion-related methods is used for prediction or decision making, whereas a content valid test describes the level of skill a person brings to the job. Both of these statements are true, of course, but I contend that we should be concerned mainly with the predictive aspects of tests used for employment decisions. We should not be primarily concerned with the descriptive aspects of inferences from a content developed test. Surely scores from a well developed typing test can be used to describe a person's skill at manipulating a typewriter, but description is not our primary purpose when we use a typing test to make hiring decisions. We would not care about the mere description of an individual's typing ability unless we were deciding to employ him or her. We essentially use the typing score to predict how successfully someone will perform a job involving typing. (Tenopyr, 1977, p. 49)

Thus, in the employment context, the association that is often drawn between the purpose of testing (i.e., prediction) and the proper design for justifying prediction (i.e., a criterion-related strategy) is a meaningless one. Furthermore, it perpetuates the stamp-collecting mentality by suggesting that if one is interested in predicting some type of future behavior, then the only appropriate analysis is embodied in the criterion-related design. A strong argument might be made that a criterion-related strategy, by itself, is rarely appropriate. In basic research, correlational results are used either to confirm an articulated a priori proposition (as in the case of using a test of a corollary to a basic deductive proposition) or to suggest a line of investigation to be pursued using a more elaborate design (commonly called the "fishing expedition"). In neither case would a correlation coefficient stand on its own. In the first case, it follows directly from a theory. In the second case, it is the precursor of theory building. What enduring value can the correlation coefficient have outside of these two contexts?

As both the Division 14 Principles (American Psychological Association, Society for Industrial and Organizational Psychology, 1980) and Tenopyr (1977) asserted, testing has no purpose if it is not used to draw an inference about future behavior (i.e., to predict). The problem seems to stem from the fact that although the purpose of employment testing is prediction, a criterion-related design is not the only, or possibly not even the best, way to support the inference that people who do better on the test will do better on the job. As suggested earlier, the strongest support for such an inference must be more than the simple statement of statistical association. Among other things, one needs to know something about the integrity of the criterion. Cronbach (1985) has argued persuasively that the determination of criterion integrity is actually an instance of construct validation. Some assurance is needed that the criterion is measurable in some reasonable way and that they have actually identified that reasonable way. Furthermore, in choosing one criterion rather than another, it is assumed that the criterion chosen is either representative of or preeminent among the many criteria that might have been chosen. This assumption implies domain sampling or what is frequently referred to as "content-oriented validity."

Lawshe (1985) provided a good example of the inadvertent linking of the purpose for testing and appropriate validity design in the following statement: "If the purpose is to infer how well a candidate will perform on the subject job (thus to predict future behavior), a criterion-related strategy is indicated" (p. 237). It is not that statements like this one are wrong; rather, they are insufficient. As illustrated earlier, having done a "good" traditional criterion-related study, the researcher or practitioner has included elements of a content and construct
strategy as well. Similarly, in carrying out a validation study using a construct model, the researcher or practitioner gives careful consideration to prior research using the construct in question. This consideration plays a substantial role in developing the nomological network—the observed pattern of relationships between the construct in question and other constructs. In employment testing, this nomological network often includes criterion-related studies carried out by other practitioners and researchers. In earlier days, one of the major personnel research journals, Personnel Psychology, even published a section called “Validity Exchange” in each issue.

Aspects of validity cannot be easily separated from one another. Because the words content, criterion-related, and construct can be used as aids in discussion, one should not be seduced into thinking of those words as standing for discrete and independent processes. Instead, the words simply represent parts of a larger system that addresses the goal of hypothesis testing. George Orwell cautioned that the worst thing that one can do with words is to simply represent parts of a larger system that addresses the goal of hypothesis testing. George Orwell cautioned that the worst thing that one can do with words is to surrender to them. The trinitarian doctrine of validity is a case of unconditional surrender. The boundaries among these three modal types of validity are a great deal more permeable than the Uniform Guidelines seem to admit. This is as it should be from the unitarian perspective. There are not different types of validity. The working hypothesis remains the same: \( Y = f(X) \). The question is what form of evidence the investigator will choose to test that hypothesis. Investigators should not lose sight of the fact that validity studies are attempts to develop a theory of performance that explains how an individual can (or will) meet the demands of a particular job.

Constructs and the Validation Process: A Special Problem

In the forensic discussions of pornography, one can often hear the observation, “I may not know how to define it, but I know it when I see it.” The implication is that there is some inchoate or primitive recognition of pornographic material and that once the material is labeled as pornographic, it is a simple matter to apply or develop laws or other administrative procedures to deal with the distribution of the material.

The same logic seems to hold in discussions of construct validity. Although there is no general agreement concerning the nature (or definition) of constructs, neither is there reluctance on the part of many individuals to brush away that conceptual problem and get on to the more practical and immediate labeling process. Consider the following observation about the propriety of a construct-related validation design: “And if we wish to infer the degree to which a candidate currently possesses a trait or other characteristic (usually a psychological construct) critical to job performance (thus to assess an attribute) a construct validity analysis is indicated” (Lawshe, 1985, p. 237). This statement represents a rather standard line of argument, one highlighted earlier in the quote from the Guardians Association (1980) appeal decision: Some variables measured by psychological tests are constructs, and others are not. If one wants to measure a construct, one should use a construct-validation strategy; if one wants to measure something else, then a criterion-related strategy may be appropriate. As Judge Newman noted, it is this decision that often determines the prevailing opinion in Title VII litigation. However, one is left wondering how the construct–nonconstruct determination will be made and by whom. The Uniform Guidelines on Employee Selection Procedures (1978) are not much help. They include no definition of a construct in either the body of the document or the glossary. Thus, without dealing directly with the most basic issue, the conceptual and operational aspects of constructs, the Uniform Guidelines encourage the “know it when I see it” attitude. The Uniform Guidelines simply identify examples of things to be considered constructs. These examples include intelligence, aptitude, personality, common sense, judgment, leadership, and spatial ability. As is explained later, if one chooses to gather predictor information on variables that can be linked in any way to the examples of constructs just listed, then construct validation seems to be required.

In all fairness, the Uniform Guidelines show justifiable concern that some psychological tests purport to measure things about which there is little conceptual or operational agreement. By extension, then, the Uniform Guidelines seem to imply that if one chooses to test for one of these poorly defined attributes, greater care is needed. Few would disagree with that interpretation. Nevertheless, in the attempt to make that simple assertion, the word construct was applied, and the seeds of confusion were sown. Labeling a variable as a construct in the context of equal employment litigation is a form of assault often chosen by plaintiffs. The label in itself may be sufficient to do the damage. If it can be established that constructs are being measured, then it seems to follow that the only appropriate validation strategy is construct validation. One is drawn back into the trinity once again. What is not obvious in this strategy is that two important assumptions are made. First, it is assumed that one is only interested in the numbers that result from the test. The second assumption is that some things that are measured are constructs and others are not. There is, however, an alternative assumption that is equally compelling and serves to focus attention back on the issue of the meaning and value of constructs in psychological measurement. In validation research, there is only superficial interest in a test score, per se. The test score is simply the result of what one is interested in—an attribute of the person taking the test (Guion, 1980). It is this attribute that is inextricably bound to the notion of a construct.

Constructs and Psychological Measurement

One line of reasoning suggests that constructs underlie all psychological measurement. Tenopyr (1977) has suggested that any inference either to prediction or to test scores involves constructs. Certainly typing ability implies an inference about an underlying characteristic on which
individuals differ. One must make assumptions about the extent to which the particular demonstration (i.e., the typing test) is representative of the individual's transcendental ability. In other words, one must assume that there is an underlying typing ability that will manifest itself similarly with a different typewriter in a different setting, with different levels of motivation and anxiety and different material to type. The performance of an applicant on a typing test is simply a point on a continuum, and that continuum is not directly observable. However, one attempts to illuminate the continuum by gathering a sample of behavior that is hypothesized to vary as a function of that underlying attribute. Similarly, in measuring something like reasoning, one must engage in certain operations to illuminate that ability continuum. As is true of typing ability, reasoning ability is not directly observable. In that sense, typing ability is similar to reasoning ability or strength or memory. None of these abilities can be directly observed. To be sure, it is easier to illuminate the typing ability or memory or strength continuum than the reasoning continuum, but all are constructs nonetheless.

Constructs and Cognitive Abilities

In the context of this discussion of construct-validation strategies, one section of the Uniform Guidelines (1978) is as destructive to the interface of psychological theory and practice as any that might have been conceived.

A selection procedure based on inferences about mental processes cannot be supported solely or primarily on the basis of content validity. Thus, a content strategy is not appropriate for demonstrating the validity of selection procedures which purport to measure traits or constructs such as intelligence, aptitude, personality, common sense, judgment, leadership or spatial ability. (14 C [1], p. 38302)

Two things seem to follow from this section of the Uniform Guidelines. First, if a variable is labeled as a construct, then the use of content-oriented validation strategies is immediately precluded. Second, mental abilities or processes are identified as vulnerable to attracting criterion-related study. The paragraph in question implies that content-oriented strategies are inappropriate for supporting inferences based on tests that measure mental processes (i.e., constructs). The logic goes something like this: Mental processes cannot, by definition, be observed. Anything that cannot be observed does not exist; it is a construction in the eye of the observer. Any such construction qualifies as a construct. Inferences based on tests that measure constructs can only be supported using a construct-validation strategy. Therefore, any test of mental processes requires construct validation.

This logic becomes more problematic when one considers the description of construct-validation strategies presented in the Uniform Guidelines: Any construct-validation effort must include a criterion-related study. This is the “catch-22” that Judge Newman so clearly identified. If one wanted to and was able to conduct a criterion-related study, why would one ever want to conduct a construct-validation study?

The dilemma results from a rigid interpretation of the phrase “conduct a criterion-related study.” The criterion-related study need not be conducted by the same individual who is providing the construct-related evidence. In addition, it seems reasonable to apply wide latitude to the notion of the criterion in such a criterion-related study. Given the context in which the word criterion is used in the Uniform Guidelines, the implication is that the criterion must be some measure of job performance or success. However, that is only one interpretation of criterion as it might be used in discussing the concept of a nomological network of relationships, a cornerstone of construct validity. I suspect that in the context of construct approaches, the term criterion-related should take on a substantially broader meaning. Instead of the typical interpretation of the term criterion as a measure of work performance, in the context of construct validation, criterion should probably be interpreted to mean some independent measurement of the attribute in question. This interpretation would allow for the traditional approaches (e.g., multitrait, multimethod designs) as well as some novel strategies for establishing the meaning of the construct. The major value of construct approach is in establishing the meaning of scores based on a test that purports to measure an attribute that is not well established in the psychological literature. A traditional criterion-related study is only one way of achieving that goal, and even the value of this traditional design will depend on how much the practitioner knows about the criterion chosen and the extent to which the predictor and criterion jointly form a conceptual theory of performance for the job in question.

In essence, the Uniform Guidelines require construct validation to reduce feelings of discomfort about the conceptual or operational meaning of the predictor. However, the narrowly defined criterion-related study that is included in the Uniform Guidelines' peculiar version of construct validation is the least likely source of information about a questionable predictor. The prototypic criterion-related study says that predictor scores are or are not related to criterion scores, not why they are related. This is the basic mechanism of covariation that has fueled most of the empirical sciences—induction. From a set of observable relationships, a rule or principle is derived. The major difference is that in the context of equal employment research, rules or principles are seldom derived. This seems to be a case of arrested induction. A narrowly conceived criterion-related-study requirement is ill suited to the intent of construct-validation strategies. However, a broadly defined criterion-related study, such as that described earlier, would be ideally suited to construct validation, because it would help to elaborate the nomological network that provides meaning to the measures.

If one does not accept the premise that some variables that are measured are constructs and others are

---

2 This notion of an expanded interpretation of criterion-related validity was suggested to me by Robert M. Guton.
not, there is no reason to consider the construct–nonconstruct question, nor is there any reason to make the content validation–construct validation distinction based on what the test measures. In short, the Uniform Guidelines ask one to accept a set of rules, premises, and assumptions that often make no sense. They ask one to join in making an arbitrary distinction that makes the rules of stamp collecting more applicable. As soon as one accepts the distinction between constructs and nonconstructs, one has no option other than to follow the checklist mentality of determining validity. Similarly, if one accepts the notion that the criterion-related evidence requirement of the construct-validity approach to validation means that the criterion must be a measure of job performance, one is no longer considering construct validation as traditionally defined. At one time or another, most of us have used the “spouse-beating” gambit (i.e., asked someone the equivalent of the question, “Do you still beat your spouse?”). The person answering is at a distinct disadvantage if he or she chooses to answer the question as asked. The same is true in the context of the construct discussion. If the distinction between constructs and nonconstructs is accepted as implied by the Uniform Guidelines, the nature of the discussion shifts almost immediately from a consideration of the adequacy of the design for supporting inferences to a two-part labeling exercise. The first part is labeling the variables of interest as constructs or nonconstructs. The second part is choosing one of the three traditional validity models to support the inferences drawn from the test scores.

The Content–Construct Continuum

Measurement theorists and practitioners have suggested that there is a continuum that might be used to distinguish between circumstances amenable to content-oriented strategies and circumstances amenable to construct strategies. I touched on this point briefly in an earlier section. Although Lawshe (1985) is the most recent theorist to repeat this assertion, the proposition itself is not new. It has been suggested by Tenopyr (1977, 1984) and is clearly spelled out in the Guardians Association appeal decision written by Judge Newman (Guardians Association v. Civil Service Commission of the City of New York, 1980). Lawshe (1985) advanced the notion of a continuum by proposing (at least by implication) three parameters to consider in choosing between content-oriented and construct-oriented validation processes. The first parameter is the extent to which the behavior in question is observable. Lawshe suggested a distinction between the observable and the reportable. Based on his examples, Lawshe implied that typing ability or arithmetic ability can be observed, but reasoning ability can only be reported. The clear implication is that if a behavior cannot be observed (but only reported), a construct-oriented strategy is in order.

A second parameter proposed by Lawshe seems to be a simple–complex parameter with “simple proficiencies” (arithmetic and typing) at one end of the continuum and cognitive processes (presumably less simple, e.g., inductive and deductive reasoning) at the other end of the continuum. Thus, he considered content-oriented strategies to be appropriate for simple proficiencies but inappropriate for complex proficiencies.

A third parameter suggested in Lawshe’s discussion is a concrete–abstract parameter. Thus, according to Lawshe (1985):

Comparison of behavior elicited by a test to behavior exercised on a job requires little or no inference at the “observation” end of the continuum; however the higher the level of abstraction, the greater is the “inferential leap” required to demonstrate and support the validity of the inference. (p. 238)

In Lawshe’s scheme, content-oriented strategies would be appropriate for concrete behavior but not abstract behavior. Note that arithmetic ability, an example of a simple and concrete attribute according to Lawshe, has not been directly observed. Instead, it is reported that the individual in question does or does not have the ability based on certain responses to number problems. With multiple-choice response formats, the practitioner is at a still greater distance from the behavior.

Lawshe (1985) concluded his discussion by warning that “content validity analysis strategies are appropriate only when the job behavior under scrutiny falls at the observation end of the continuum; when such behavior approaches the abstract end of the continuum, a construct validity analysis strategy is indicated” (p. 238). Once again, there appears to be a backsliding into the trinity of criterion, content, and construct models of validation, and once again, constructs appear to be at the heart of the decision as to which model is appropriate.

It may be of greater value to focus on the meaning and role of constructs in test development and employment decision making than on the extent to which one validation model or another is most appropriate. I believe that constructs are central to all of psychology. Other sciences have not been unduly hampered by the adoption of constructs as both descriptive and explanatory mechanisms. The periodic charts of elements, gravity, electricity, cancer, and socioeconomic status are all constructs that have served a useful purpose at one time or another in the scientific endeavor. In employment testing, strength, typing, arithmetic abilities, and memory have all played substantial roles in the prediction of employee success.

The fact that motor abilities, such as strength, or cognitive abilities, such as memory, are accepted without discussion as observable illustrates the willingness of the stamp collectors to strain at gnats while swallowing camels. Having a candidate identify an object, person, or phrase that was encountered 10 minutes (or 10 days) previously is no more or less observable than asking a candidate to apply a general rule to a specific case and choose a response from a set of responses that best represents the application of the rule (i.e., to reason deductively). Similarly, few would argue with the proposition that inferences drawn from knowledge tests can be supported using content-oriented strategies (assuming, of course, that the knowledge can be shown to be critical to the job and not de-
veloped in a training program subsequent to appointment). However, this proposition begs the question of the observability of the knowledge. Does choosing the correct answer make the knowledge observable? If so, then the same operation renders memory, reasoning, verbal comprehension, and other cognitive abilities observable. Similarly, how might one distinguish between the possession of a body of information (knowledge) and the present competence to retrieve information previously encountered (memory)?

**Conclusions**

Applied measurement specialists have only recently and somewhat reluctantly accepted the cognitive revolution that altered all of scientific psychology in the late 1960s. In the area of testing, there seems to be a determination to maintain a distinction between behavior and mental processes. With respect to equal employment legislation and litigation, they are encouraged to continue this separate (and unequal) status by a body of administrative law (the Uniform Guidelines) developed for the convenience of those who prefer stamp collecting to scientific investigation. It is clear that the status and definition of constructs are sources of continuing debate. MacCorquodale and Meehl (1948), Margenau (1950), and Bechtoldt (1959) were pioneers in suggesting the role of constructs in the scientific enterprise. More recently, Cronbach (1971, 1985) and Stenner, Smith, and Burdick (1983) suggested that circumstances can be created that permit the "observability" of constructs. Whitty (1983) proposed that the recent emphasis on information-processing models of human behavior has greatly enhanced the value of constructs in psychological measurement. This is no time to be letting the administrative tail (the Uniform Guidelines) wag the psychological dog (the role of constructs in measurement).

If applied psychologists and measurement specialists are to bring anything of value to the fair employment arena, they should not be persuaded to leave their strengths in the locker room. As Lawshe (1985) reasonably argued, psychologists are good at designing studies that conform to the generally accepted model of scientific hypothesis testing. They should not be persuaded to accept the limited concepts of deductively oriented practitioners (personnel administrators, lawyers, and judges) if accepting those concepts suspends the growth of and diminishes the utility of their own profession. From that perspective, it is acceptable to consider the content, criterion-related, and construct validation strategies as generally descriptive of how one might go about supporting any hypothesis, including one that examines the relationship between test performance and job performance. Nevertheless, psychologists should not aid and abet the stamp collectors who are determined to identify the right type of approach (or type of analysis) for every situation. Furthermore, psychologists should put an end to the artificial distinction between behavior and mental processes, a distinction that the Uniform Guidelines seek to maintain. Instead, they should begin to apply the traditional standards for evaluating experimental design and hypothesis testing to all efforts, both conceptual and empirical, that seek to support inferences from test scores. In the most basic sense, psychologists must stop thinking of testing and the validation process as something different from traditional hypothesis testing.

If the validation process is accepted as a form of hypothesis testing, then the goal becomes one of collecting the best, most informative, and least ambiguous evidence that resources allow. The better the evidence, the more alternative explanations for the relationship between test performance and job performance that can be ruled out. Guion (1984) has suggested recently that the concern should be with the "preponderance of evidence" in drawing inferences rather than with the type of validation model that is used. This way of thinking should not be uncommon to a judge or a lawyer. To be sure, three models of validity would be easier to deal with, because they would allow for the development of checklists. Nevertheless, in other forensic activities, the "weight of evidence" model is widely accepted.

Psychologists understand how to design studies that can answer questions about behavior. They understand how to rule out alternative explanatory mechanisms. In addition, they know how to construct theories of behavior based on the results of those studies. These are their strengths, and psychologists should not implicitly accept a set of rules and procedures that obviate these strengths. It is their creativity in design that provides psychologists with legitimate and expert power. Allowing their capacity to design studies and collect evidence to be limited by administrative practice (e.g., acceptance of only three models of the validation process) diminishes the quality of the evidence they can gather. If psychologists continue to accept the tripartite division of validation models suggested in the Uniform Guidelines, they will become pitiful giants wandering aimlessly on a landscape being inexorably devastated by those with little interest in or understanding of human behavior and its prediction. By virtue of education, experience and research, psychologists are well prepared to render expert opinion about validity designs and inferences based on test scores. The stamp collectors would mock and limit that right and responsibility.

**REFERENCES**


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


