4. The Status of Test Validation Research

Lyle F. Schoenfeldt
Texas A&M University

Follow this and additional works at: http://digitalcommons.unl.edu/burostestingissues

Part of the Educational Assessment, Evaluation, and Research Commons

http://digitalcommons.unl.edu/burostestingissues/6

This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Social and Technical Issues in Testing: Implications for Test Construction and Usage by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
The Status of Test Validation Research

Lyle F. Schoenfeldt
Texas A & M University

More than any other area, validation research is where the “rubber meets the road” in test construction and test usage. The very term validation implies the assessment or measurement of individuals and the relationship of this assessment to some criterion of performance. The success of a test validation effort, or the lack thereof, has implications for the value of the assessment and for the utility of the procedures.

In today’s environment, whether the validation is intended for employee selection, educational decisions, or personal counseling, there is an increasing probability that the outcomes of research will have legal implications. In the past, a testing program could be set up in terms of professional judgment without including the experimental validation of the procedures. If the individuals involved in establishing the test program were knowledgeable, it was quite possible the tests, although unvalidated, would make a practical contribution in terms of the goals intended. In the absence of a formal validation, however, one would never know the extent to which the testing program was successful or superior to another assessment procedure. A testing program that does not involve validation research is at best an unknown and at worst may be an outright fraud. In either case, the likelihood that testing procedures will have to be defended, including the possibility of legal action, has increased dramatically.

The purpose of the present review is to look carefully at the current status and future directions of test validation research. It will be of value to look at what we know, some of the problems with the process by which tests have been validated up to now, what needs to be learned, and how we will move ahead in the area of test validation research. Finally, it will be important to consider test validation research as a vehicle for improving test construction and test usage.
Test Validation: A Definition

In the context of this discussion a test is defined as any measure, combination of measures, or procedure used to evaluate differences among people. In this manner, the term tests includes the full range of assessment techniques from traditional paper and pencil tests to performance assessments, and includes such things as training programs (e.g., school achievement), situational assessment, and probationary tryouts. In other words, a test is any formal or informal assessment from which an inference is drawn. For example, if a student transferring into a middle school were to be given a series of paper and pencil assessments as a basis for determining course assignments, few would disagree that these assessments constitute a test. On the other hand, the same decisions could be made on the basis of an interview between a school counselor, the student, and parents. Because inferences about readiness for various courses result from the counselor–student interaction, one could consider that this is also a test.

Validity is the degree to which inferences from scores on tests or other assessments are supported or justified on the basis of actual evidence. Validity is not a characteristic of a test; rather it is a characteristic of inferences that result from a test, assessment, or observation. Thus, validation determines the degree of relatedness between inferences made and actual events.

History of Test Validation

The history of measurement and validation is at least as old as Plato’s Republic. Various summaries of the important events surrounding modern mental measurements have been well documented (Linden & Linden, 1968). In his review of the role of tests in personnel selection, Guion (1976) developed a series of tenets that summarize the “orthodox” history of validation research. These tenets, as adapted from Guion (1976), are summarized in Table 4.1. As seen, the emphasis is on developing a singular predictor–criterion relationship as the basis for determining validity. The dates in the table suggest that the tenets were well established early in this century. Further, these values would not be wide of the mark in the 1980s for an investigator interested in a traditional validation project.

TRADITIONAL APPROACHES TO TEST VALIDATION

Criterion Related

Traditionally, the criterion-related approach has dominated validation research. The “tenets” of criterion-related research are essentially those described by Guion (1976) and summarized in Table 4.1. It is possible to distinguish two alternate approaches within the criterion-related procedure. Concurrent validation involves the relationship of tests to criterion measures obtained at the same
## TABLE 4.1
Guion's Historical Tenets of Orthodox Validation Research

<table>
<thead>
<tr>
<th>Tenet</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The purpose of validation is to predict future performance.</td>
<td>“It is . . . essential to know whether the scores are in any useful sense predictive of subsequent success [Bingham, 1937, p. 216].”</td>
</tr>
<tr>
<td>2. Predictors and criteria should be selected on the basis of job analysis.</td>
<td>“the tests which are to be experimented with can be chosen only on the basis of some more or less plausible relationship between particular tests and the sort of duties performed [Kornhauser &amp; Kingsbury, 1924, p. 47].”</td>
</tr>
<tr>
<td>3. Measuring instruments must be standardized.</td>
<td>“In order for measurements of persons taken at varying times to be comparable, the procedure of the test must be uniform [Freyd, 1923, p. 232].”</td>
</tr>
<tr>
<td>4. Tests should be empirically validated.</td>
<td>No test has any significance before it is tried out (Link, 1924).</td>
</tr>
<tr>
<td>5. Validation is situation-specific.</td>
<td>“if maximum value is to be attached to test scores the conditions under which the . . . [examinees performed] with the use of tests should reproduce in general the conditions under which they . . . [performed] when the tests were evaluated [Freyd, 1923, p. 381].”</td>
</tr>
<tr>
<td>6. More than one test should be used.</td>
<td>To quote Guion (1976, p. 783), “Hull (1928) insisted that a battery of four or five tests or more must be developed if the criterion in all its complexity was to be predicted with maximum efficiency.”</td>
</tr>
<tr>
<td>7. Only one criterion should be used.</td>
<td>Freyd (1923, p. 223) described the process by which “a criterion” should be selected.</td>
</tr>
<tr>
<td>8. Tests are preferred over non-test assessments.</td>
<td>“The experimenter will not limit himself to any particular type of measuring instrument, but those in which he will be most interested are tests and questionnaires [Freyd, 1923, p. 231].”</td>
</tr>
<tr>
<td>9. Individual differences should be recognized in evaluating tests.</td>
<td>“If men and women are both . . . [involved in the validation research] it will be necessary to examine the results for sex differences, and if need be, to evaluate the test separately for the two sexes [Freyd, 1923, p. 225].”</td>
</tr>
</tbody>
</table>

*aAdapted from Guion (1976). Copyright © 1976 by John Wiley & Sons, Inc. Used by permission.*
time as the test data. Predictive validity involves the assessment of individuals followed by the collection of criterion information at some subsequent time. In some designs, the time factor can be an important consideration, whereas in other situations it is not. For example, in predicting job success, concurrent validation inevitably involves existing employees whose motives for performing well on the test may differ from the motivation of applicants. In other fields, such as psychometrics, concurrent validity is used to demonstrate, for example, that a paper and pencil assessment is an adequate substitution for a more cumbersome, painful, or inefficient assessment procedure. In both cases, though, the goal is to develop and to test a hypothesis and (hopefully) to assert validity on the bases of a demonstrated relationship between individual characteristics and measures of performance.

Criterion-related validity has traditionally been the most frequently used approach to test validation. In any instance of criterion-related validity, most attention is usually given to the decision about the selection of the criterion variables. Given that the validation process is one of inferences from test scores, the definition of the criterion or standard to be inferred looms large as a possible limitation in the criterion-related approach.

The fact that two relatively recent review articles dealt with this subject (James, 1973; Smith, 1976) emphasizes the attention that criteria selection is receiving. Although the orthodox tenets of the traditional approach focus on a single criterion, which often is a weighted combination of several criteria or a succession of single measures, the emphasis of these two reviews is on a more complex approach to the development of criteria. Mention is made in these reviews of various models including the ultimate criterion (Thorndike, 1949), the complete final goal of a particular type of selection or training; multiple criteria approaches (Dunnette, 1963; Ghiselli, 1956; Guion, 1961; Wallace, 1965) (as exemplified by the model shown in Fig. 4.1 and discussed later); and general criterion models (as exemplified by the models shown in Figs. 4.2 and 4.3 and discussed in a later section).

Content Oriented

Another traditional approach to the validation of tests is the content-oriented procedure. This approach is applicable when empirical investigation is not possible and involves validation on the basis of assumed or hypothesized relationships. The legitimacy of the content-oriented procedure lies in the degree to which the hypothesis itself is well grounded in carefully controlled observations and prior research results (Guion, 1976). Although mentioned in various texts and in the Standards for Educational and Psychological Tests (American Psychological Association, 1974), content-oriented validation has always been the stepchild of testing. Until quite recently information about procedures for demonstrating content-oriented validity has been perfunctory, contradictory, or un-
TABLE 4.2
Steps in Content Validation

1. Task analysis
2. Definition of performance domain
3. Survey of performance domain
4. Development of items
5. Demonstration that items constructed are representative of the performance domain
6. Development of cut-off score

available. The emergence of content-oriented validity has been largely a result of a series of conferences (Guion, 1974a; Proceedings, 1975), articles (Guion, 1974b, 1977; Schoenfeldt, Schoenfeldt, Acker, & Pearlson, 1976; Tenopyr, 1977), and manuals (American Psychological Association, 1974, 1975, 1980; Mussio & Smith, 1973). The steps involved in a study of content-oriented validity are summarized in Table 4.2.

Perhaps the criticism of these two approaches to validation has been best exemplified by Loevinger’s (1957) belief that criterion-related validities are “ad hoc” and that content-oriented validity relies too much on the judgment of the investigator and is thus nongeneralizable. Loevinger believes that ad hoc arguments are scientifically of minor importance if not actually inadmissible and terms both approaches to validation as “administrative” as her way of implying a lack of scientific basis.

CONSTRUCT VALIDATION

Definition of Construct Validity

Construct validity is concerned with understanding the underlying dimensions or attributes being measured through any test or observation process. This type of validation is less concerned with specific performance inferences but instead considers the relationship of test scores to possible underlying attributes.

Many researchers have conducted validation studies but tend to show little concern for construct validity. Construct validity is more in the nature of determining the scientific basis of a particular measure and frequently does not concern practitioners. Evidence of construct validity is often found in a well-developed manual accompanying a particular test or is obtained by pulling together the results of studies dealing with a particular instrument. With regard to the latter, The Eighth Mental Measurements Yearbook (Buros, 1978) lists over 5000 references to the Minnesota Multiphasic Personality Inventory (MMPI). Undoubtedly, the totality of this massive body of research provides much valuable information about relationships to other tests, to criteria, and (through various multivariate analytic procedures) to numerous constructs.
On the basis of relating particular measures to a wide variety of possible performance outcomes or other test scores, a network of research data is developed from which inferences could be drawn about the nature of the original test and the constructs that underlie it. Large-scale studies of construct validity are done and form the basis for new scientific learning about specific measures in particular and human differences in general. More than with other approaches to validation, a successful study of construct validity suggests and encourages further research.

History of Construct Validity

Construct validation has always existed, at least at an implicit level, but was only formally defined and extensively discussed in the mid- to late 1950s. A quote from Cronbach and Meehl (1955) best summarizes the early articulation of this conceptualization:

Validation of psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950–54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee’s report was the term construct validity. This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee . . . . The statements agreed upon by the Committee (and by committee of two other associations) were published in the Technical Recommendations . . . . Identification of construct validity was not an isolated development. Writers on validity during the preceding decade had shown a great deal of dissatisfaction with conventional notions of validity, and introduced new terms and ideas, but the resulting aggregation of types of validity seems only to have stirred the muddy waters. Portions of the distinctions we shall discuss are implicit in Jenkins’ paper, “Validity for what?” (1946), Gulliksen’s “Intrinsic validity” (1950), Goodenough’s distinction between tests as “signs” and “samples” (1950), Cronbach’s separation of “logical” and “empirical” validity (1949), Guilford’s “factorial validity” (1946), and Mosier’s papers on “face validity” and “validity generalization” (1947, 1951). Helen Peak (1953) comes close to an explicit statement of construct validity as we shall present it [p. 281].

Further discussions by Loevinger (1957), Bechtoldt (1959), Campbell (1960), and Ebel (1961) followed, and all contributed in refining the definition of construct validity as well as in compiling evidence necessary to substantiate its existence.
Multitrait–Multimethod Approach

In terms of providing a methodology to verify construct validity, the article with by far the greatest impact was “Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix” by Campbell and Fiske (1959). In this seminal work, Campbell and Fiske (1959) advocated a procedure for triangulating a construct, utilizing a matrix of intercorrelations among tests representing at least two traits, each measured by at least two methods. Construct validity is the degree to which measures of the same trait correlate higher with each other than they do with measures of different traits involving separate methods.

The importance of the multitrait–multimethod (MTMM) procedure is in the provision of a conceptualization of construct validity that could be readily operationalized by researchers. Interestingly, few articles or dissertations were published in the 1960s using the MTMM approach. The rate of diffusion of the technology was understandably slow. However, the MTMM procedure has come into its own in the 1970s and 1980s. An extensive computer review of the validity literature revealed that 10 articles/dissertations were published in 1979 and another 12 were published in 1980, using the MTMM approach. This is exemplary of how standard the procedure has become in the establishment of construct validity.

There have been both extensions and critiques of the MTMM. Werts, Joreskog, and Linn (1972) suggested that the MTMM approach may be treated as a problem in confirmatory factor analysis and that the MTMM is subsumed by the general model for analysis of covariance structure. Other authors have proposed further innovative factor analytic applications (Golding & Seidman, 1974; Jackson, 1975; Kenny, 1976; Levin, 1974; Ray & Heeler, 1975). Other extensions have been in the application of nonparametric statistics (Hubert & Baker, 1978) and path analytic procedures (Schmidt, 1978). Limitations of the MTMM have been discussed by Kalleberg and Kluegel (1975).

Other Approaches to Construct Validity

The multitrait–multimethod procedure has clearly become a standard for the establishment of construct validity. At the same time, given the definition of construct validity discussed previously, it is obvious that researchers are not limited in the number of procedures employed to establish its existence. In fact, given the nature of content validity, it is somewhat heretical to focus on methods rather than models, although to a large extent the two are closely linked in the context of this topic.

Historically, factor analysis has been associated with the establishment of constructs. Many applications of factor analysis are in the nature of data reduction, and as such the results have little in the way of implications for the establishment of construct validity. However, in conjunction with an appropriate
model, factor analysis can play a valuable role in the validation of constructs. Guilford's (1967; Guilford & Hoepfner, 1971) extensive work on the structure of intellect is one of many examples that could be cited illustrating how a model and appropriate factor analytic procedures can come together in the establishment of construct validity.

Another method receiving recent recognition as a vehicle for its contribution to the establishment of construct validity are latent-trait models (LTM). Several recent studies by Whitely (1980a, 1980b) provide an example of the potential contribution of LTM to the study of intelligence. LTM resolve several measurement problems in studies of intellectual change, including ability modification and life-span development. LTM contribute to construct validity in their capability to represent an individual differences model of cognitive processing on ability test items.

**Construct Validity: State of the Science**

Although specific procedures play an important role in the demonstration of construct validity, the more important priority should be the research design. With regard to the latter, some of the most recent work was discussed in a conference on *Construct Validity in Psychological Measurement* (U.S. Office of Personnel Management, 1980). This conference involved several important themes. First was a call for more clearly defined professional standards for construct validity. Second was a discussion of the realization of the role construct validity plays, in conjunction with criterion-related and content-oriented validity, in the assessment of human differences. Included in this theme was the singularly unique application of a construct model in the validation of the Federal Government’s Professional and Administrative Career Examination (PACE), as reported by McKillip and Wing (1980).

A third theme of the conference involved a review of thinking and progress in several important areas of assessment by several recent contributors in each area. Carroll (1980) discussed background and progress in his assessment of abilities. Sternberg examined different approaches to the construct validity of aptitude tests in the context of an information-processing assessment (Sternberg, 1980). Jackson (1980) reviewed construct validity and personality assessment, concluding “that through a judicious combination of psychological analysis of dispositional variables and psychometric and multivariate procedures, progress in personality assessment is possible [p. 79].” Frederiksen (1980) and Messick (1980), in different presentations, discussed research models for construct validation.

In his conference review, Dunnette (1980) developed a number of integrating thoughts with respect to construct validity. One of his main points was that, as a part of a scientific undertaking, the study of constructs should be pursued by diverse research strategies. Certainly anyone present at the conference or familiar
with the proceedings would be impressed with the diversity of approaches taken and with the state of the art with respect to scientific knowledge about intelligence, aptitude, and personality constructs.

**MULTIVARIATE VALIDATION MODELS**

Psychologists and measurement specialists have been interested in predicting human behavior over a long period of time, although the shape and form of this interest has changed. Traditional interest was largely empirical and has been based on linear methods of prediction. Typical results have been disappointing. For example, Ghiselli (1966) has summarized 107 validity coefficients calculated to predict training and proficiency criteria. The mean validity coefficients in five major aptitude areas are shown in Table 4.3. As seen, coefficients are relatively modest, with the overall average correlation to predict training success being .30 and to predict the more important criterion of job performance, .19. These results have spurred many researchers to experiment with various multivariate models over the last 15 years.

**Person–Process–Product Models**

One class of approaches might be termed *person–process–product* models in that they attempt to examine behavior as a complex outcome of interactions between individual attributes and organizational requirements within the setting in which the behavior occurs. Figure 4.1 is a schematic portrayal of a prediction model adapted from one suggested by Guetzkow and Ferehand (1961). It was designed in an effort to take into account complex interactions that may occur among various predictor combinations, different groups or types of individuals, different behaviors, and the consequences of these behaviors. As Dunnette (1963) indicated, the model permits the possibility of predictors being differentially useful for predicting the behaviors of different subsets of individuals. Also evident is the fact that similar behaviors may be predictable by different patterns of interaction between groupings of predictors and individuals or even that the same level of performance on predictors can lead to substantially different patterns of behavior for different people. Also, incorporated into the model is the fact that the same or similar behaviors can lead to quite different outcomes depending on the situation.

A similar model, couched in terms of predicting job performance, is shown in Fig. 4.2 (Campbell, Dunnette, Lawler, & Weick, 1970). In this model, job performance is viewed as a product of the person impacting with various organizational forces. The individual is represented as a configuration of abilities, special skills, interest, personality traits, attitudes, expectancies, and reward preferences.
### TABLE 4.3
Comparison of Validity Coefficients for Training and Proficiency Criteria by Type of Test

<table>
<thead>
<tr>
<th></th>
<th>Mean Validity Coefficient</th>
<th>No. Pairs of Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train.</td>
<td>Prof.</td>
</tr>
<tr>
<td>Intellectual abilities</td>
<td>.35</td>
<td>.19</td>
</tr>
<tr>
<td>Intelligence</td>
<td>.34</td>
<td>.21</td>
</tr>
<tr>
<td>Immediate memory</td>
<td>.23</td>
<td>.15</td>
</tr>
<tr>
<td>Substitution</td>
<td>.27</td>
<td>.23</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>.42</td>
<td>.15</td>
</tr>
<tr>
<td>Spatial and mechanical abilities</td>
<td>.36</td>
<td>.20</td>
</tr>
<tr>
<td>Spatial relations</td>
<td>.38</td>
<td>.19</td>
</tr>
<tr>
<td>Location</td>
<td>.24</td>
<td>.17</td>
</tr>
<tr>
<td>Mechanical principles</td>
<td>.41</td>
<td>.24</td>
</tr>
<tr>
<td>Perceptual accuracy</td>
<td>.26</td>
<td>.23</td>
</tr>
<tr>
<td>Number comparison</td>
<td>.25</td>
<td>.24</td>
</tr>
<tr>
<td>Name comparison</td>
<td>.24</td>
<td>.29</td>
</tr>
<tr>
<td>Cancellation</td>
<td>.58</td>
<td>.19</td>
</tr>
<tr>
<td>Pursuit</td>
<td>.18</td>
<td>.17</td>
</tr>
<tr>
<td>Perceptual speed</td>
<td>.30</td>
<td>.27</td>
</tr>
<tr>
<td>Motor abilities</td>
<td>.18</td>
<td>.17</td>
</tr>
<tr>
<td>Tracing</td>
<td>.18</td>
<td>.15</td>
</tr>
<tr>
<td>Tapping</td>
<td>.15</td>
<td>.13</td>
</tr>
<tr>
<td>Dotting</td>
<td>.15</td>
<td>.14</td>
</tr>
<tr>
<td>Finger dexterity</td>
<td>.16</td>
<td>.20</td>
</tr>
<tr>
<td>Hand dexterity</td>
<td>.24</td>
<td>.22</td>
</tr>
<tr>
<td>Arm dexterity</td>
<td>.54</td>
<td>.24</td>
</tr>
<tr>
<td>Personality traits</td>
<td>.05</td>
<td>.08</td>
</tr>
<tr>
<td>Interest</td>
<td>.05</td>
<td>.08</td>
</tr>
<tr>
<td>All tests</td>
<td>.30</td>
<td>.19</td>
</tr>
</tbody>
</table>

*From Ghiselli, 1966.*

Looking at the model from the individual’s point of view, a job involves task demands that are objective lists of expectancies or priorities imposed upon the individual in an attempt to alter behavior in specified ways. Due to this, an individual’s behavior consists entirely of emitted responses and performance on the job that includes those aspects of behavior related to organizational climate. The result or product of the individual’s effort is a contribution to the organization, the generalized result of performance.

The models shown in Figs. 4.1 and 4.2 are two of several that summarized the relationship between individual characteristics and outcomes. The implications are significant. Behavior is seen as a complex product of cognitive, noncognitive (including motivational tendencies), and stylistic abilities. Expenditures of ener-
Predictors and Predictor Combinations  | Groups or Types of Individuals  | Criterion Behaviors  | Situations  | Consequences
---|---|---|---|---
P_1  \rightarrow  \{i_1, i_2\}  \rightarrow  B_1  \rightarrow  | S_1  \rightarrow  C_1
P_2  \rightarrow  \{i_3, i_4\}  \rightarrow  B_2  \rightarrow  | S_2  \rightarrow  C_2
P_3  \rightarrow  \{i_5, i_6\}  \rightarrow  B_3  \rightarrow  | S_3  \rightarrow  C_3
\vdots
P_n  \rightarrow  i_n  \rightarrow  B_n  \rightarrow  | S_n  \rightarrow  C_n

FIG. 4.1. A modified model for selection and prediction (adapted from Dunnette, 1963, p. 319).

FIG. 4.2. Model for the prediction of job effectiveness (from Campbell, Dunnette, Lawler, and Weick, 1970, p. 475).
Motivation are the product of motivational forces. The level of motivation determines whether goal-oriented behavior occurs or not. Once an individual is motivated, the effectiveness of performance is determined by the cognitive capabilities, stylistic tendencies, and other attributes of the individual.

Moderator Validation

A study by Berdie (1961) suggested that persons differing in intraindividual trait variation (on measures of mathematics proficiency) might be differentially predicted to be successful or unsuccessful in engineering studies. Thus, intraindividual trait variation was thought to "modify" performance predictions. Other efforts to discover moderators in predictions were given in studies by Fiske (1957) and Fiske and Rice (1955), both of which were similar to the Berdie (1961) study. In addition, studies by Cleary (1966), Frederiksen and Melville (1954), Ghiselli (1956, 1960a, 1960b), Lee (1961), and Rock (1969) are relevant. In each case, the dominate theme has been an effort to identify persons who are consistently more (or less) predictable using particular sets of predictors or subgroups of persons requiring different prediction procedures.

The procedures described are statistical in that they all involve variations of frequently employed prediction procedures. Although some of the procedures are more difficult to implement than others, unlike the models shown in Figs. 4.1 and 4.2, all have been attempted in one or several studies.

Recently it has become apparent that moderated prediction approaches are not much better than traditional linear methods of prediction. Zedeck (1971), for example, showed that initially favorable results usually failed to maintain their superiority upon cross validation. In discussing such statistical strategies, Dunnette and Borman (1979) concluded that:

Selection research must devote increased effort toward reducing sources of both variable error (measurement and sampling error) and constant error (such as perceptual biases) in the development of instruments and in the design of studies. Non-linear models may some day once again warrant attention but not until such errors have been reduced significantly to overcome the inherently superior robustness of the simple linear model [p. 495].

Aptitude by Treatment Interactions

In 1957, Cronbach wrote of "The Two Disciplines of Scientific Psychology," the one concerned with correlation and the other, through experimentation, with the sequence of events. General discussions of the importance of combining the

---

1 Reproduced with permission, from the Annual Review of Psychology, Volume 30. © 1978 by Annual Reviews, Inc.
"two disciplines," as Cronbach (1957) has been recommending, have been published by, among others, Owens (1968, 1971) and Vale and Vale (1969). More recently Cronbach (1975) and Cronbach and Snow (1977) have published comprehensive and penetrating reviews examining the background into the nature of the problem as well as the rationale for the aptitude by treatment (ATI) procedure they advocated as an alternate validation model for enhanced prediction.

The results of the ATI approach to date have not been impressive. Evidence for significant interactions is scarce and fragmentary. Second- or third-level interactions tend to cloud any simple person–performance relationships, or at least render relationships inconsistent from sample to sample. In Cronbach’s (1975) words:

The line of investigation I advocated in 1957 no longer seems sufficient. Interactions are not confined to the first order; the dimensions of the situation and the person enter into complex interactions. . . . Taking stock today, I think most of us judge theoretical progress to have been disappointing [p. 116].

Later in the same article, Cronbach (1975) states:

When ATIs are present, a general statement about a treatment effect is misleading because the effect will come or go depending on the kind of person treated. When ATIs are present, a generalization about aptitude is an uncertain basis for prediction because the regression slope will depend on the treatment chosen. . . . An ATI result can be taken as a general conclusion only if it is not in turn moderated by further variables. If Aptitude × Treatment × Sex interact, for example, then the Aptitude × Treatment effect does not tell the story. Once we attend to interactions, we enter a hall of mirrors that extends to infinity. However far we carry our analysis—to third order or fifth order or any other—untested interactions of a still higher order can be envisioned (emphasis added) [p. 199].

Thus, in Cronbach’s own words, the ATI path he has walked in an effort to infer future performance better has not been fruitful. Gains were made, as reported in the 1975 publication, but these were of less magnitude than had been hoped might materialize. These reservations have led Cronbach (1975) to propose abandonment of the ATI approach as a potential explanatory model for predicting performance behavior.

Assessment–Classification Model

Although the list of approaches that have been attempted to improve the inferential or validation process could extend ad infinitum, one further procedure, namely the Assessment–Classification model described by Schoenfeldt (1974), is worthy of mention. The Dunnette (1963) model, and virtually all the approaches discussed in this section, sought to improve the quality of inferences
made on the basis of the assessment data by identifying subsets of persons for whom predictors were differentially useful, for whom situational factors varied, and so forth. On the basis of these concerns, as well as in the interest of an alternative to the ATI model, Owens (1968, 1971) suggested his development-integrative model. The Assessment–Classification model, shown in Fig. 4.3, is the logical extension of the Owens’ developmental–integrative approach.
and the version most compatible with the models shown in Figs. 4.1 and 4.2. Thus, it incorporates the evaluation of person, process, and product (as suggested by the models in Figs. 4.1 and 4.2) with the subgroup conceptualization formulated by Owens (1968).

The specific process involved in actualizing the Assessment–Classification model consists of providing separate categorizations of the predictor and criteria sets. In dealing with the predictor set, two steps are needed. The first step involves identifying standard predictors found to be related logically to the criteria in question. The individual differences variables of the Campbell et al. (1970) model provide an example of predictor variables that might be used. The second step requires implementation of the procedures described by Owens (1968), that is, formulating subgroups with respect to the major dimensions of antecedent behavior and relating the subgroups to relevant criteria. This entails administering a background questionnaire to assess the antecedent behaviors. On the basis of responses to this questionnaire, individuals would then be classified into subgroups that are homogenous with respect to important dimensions of life behavior. In other words, the subgroups are constructed on the basis of bringing together individuals who have reported similar background patterns.

The other aspect of the Assessment–Classification model concerns the structuring of the criteria domain, the jobs (in the case of Fig. 4.3), but with other criteria for other situations. In Fig. 4.3, the structuring of jobs into families homogeneous with respect to their performance requirements and desirable configurations of attributes is illustrated. Also, several instruments have been developed and found to be of use for measuring or structuring jobs in terms of the psychological demands required for successful performance (Cunningham, 1969; McCormick, Jeanneret, & Mecham, 1969). Other procedures would be used for structuring the criteria domain in educational or clinical settings.

Unlike the conceptual models in Figs. 4.1 and 4.2 that do not lend themselves to statistical evaluation or the statistical models that have been tried and found lacking, the results with the Assessment–Classification model have been positive. Schoenfledt (1974) examined the validity of the model with a large sample of students \( N = 1934 \) working toward college degrees. Subgroups formed on the basis of previous behavioral data collected during the freshman year differed with respect to criterion (major, grade point average, and so forth) measurements taken 4 years later. More importantly, the subgroups differed with respect to the curricular paths taken during college. The result indicated that it was possible to differentiate people in meaningful ways (i.e., to subgroup individuals and to match these subgroups with similar structuring of the criteria domain).

Two industrial applications using the Assessment–Classification model have been reported. In the first, Morrison (1977) tested the model’s efficacy in making placement decisions in an industrial setting with nonexempt employees. Eight developmental-interest dimensions describing life choices, values, and interests
of 438 blue-collar workers were formulated. Job analysis identified two clusters of positions that were homogenous within and differentiated between each other on relevant job attributes. One cluster composed of 102 incumbents with more than 6 months service consisted of process operator positions. The other cluster was composed of heavy equipment operator positions and had 148 incumbents. A discriminate function was calculated on a validation group of incumbents in an effort to develop a linear combination of the life history factors that maximally differentiated the two job families. Cross validation demonstrated that three psychologically meaningful dimensions discriminated among the groups at both statistical and practical levels of significance. The process operators were more likely to be raised in an urban environment, to have a more favorable self-image, and to prefer standardized work schedules than the heavy equipment operators.

The second study was by Brush and Owens (1979) and utilized a total of 1987 nonexempt employees of a U.S. oil company. Each employee completed an extensive biographical inventory. Hierarchical clustering of the resulting biographical profiles produced 18 subgroups of employees such that, within any one subgroup, background experience and interest were similar, and yet among subgroups they were different. A similar methodology was applied to job analysis data in creating a structure of 19 job families for 939 office and clerical jobs. Significant relationships were found between biodata subgroups and other variables, such as sex, educational level, termination rate, job classification, and (most important) performance rating.

VALIDITY GENERALIZATION

One of the tenets of the traditional criterion-related validity model has been belief in the situational nature of the results. For more than 50 years, researchers have believed that the results of criterion-related validity studies were applicable only to the situation on which the study was based. This is understandable because research, such as that by Ghiselli (1966), has clearly demonstrated results of using the same predictors to predict similar criteria using different subjects in comparable (different) settings varied over a wide range. The empirical results of Ghiselli (1966) demonstrated considerable variability in validity coefficients even when predictors and criteria were essentially identical.

On the basis of findings by Ghiselli (1966) and other investigators over a long period of years, the profession has concluded that validity generalization was essentially impossible (Ghiselli, 1966, p. 228; Guion, 1965, p. 126). This conclusion even has been incorporated into professional standards (American Psychological Association, 1975) and government regulations (U.S. Equal Opportunity Commission, 1978). In fact, Guion (1976) indicated that the problem of limited validity generalization was perhaps the most serious limitation of personnel psychology.
Bayesian Validity Generalization

Change in the belief of limited generalizability was seen in the mid-1970s and the years followed through the work of Schmidt and Hunter along with their colleagues. The initial article by Schmidt, Hunter, and Urry (1976) attacked the problem of small numbers typically used in validity studies. As pointed out in the Schmidt et al. (1976) article, it typically has been believed that sample sizes of 30 to 50 individuals were adequate to make criterion-related validity studies technically feasible. To quote Schmidt and Hunter (1980):

> When sample sizes are in the 30–50 range statistical power is typically in the .25 to .50 range. That is, if the test is in fact valid, such studies will correctly detect the validity only 25–50% of the time. Sample sizes required to produce statistical power of .90 are much larger, often ranging above 200 or 300 [p. 43].

In further articles, Schmidt and Hunter (1977) and Schmidt, Hunter, Pearlman, and Shane (1979, pp. 260–261) identified seven artifactual sources that would explain the fact that different validity coefficients would result when identical predictors and criteria were studied within the context of the same job. The seven sources of variance that might lead to different results were as follows:

1. Differences between studies in criterion reliability.
2. Differences between studies in test reliability.
3. Differences between studies in range restriction.
4. Sampling error (i.e., variance due to \( N < \infty \)).
5. Differences between studies in amount and kind of criterion.
6. Computational and typographical error.
7. Slight differences in factor structure between tasks of a given type (e.g., arithmetic reasoning test).

Schmidt, Hunter, Pearlman, and Shane (1979) proposed that a researcher with, say, 100 validity coefficients relating tests of perceptual speed to clerical proficiency compute the variance of the validity coefficient distribution and subtract variance due to each of the artifactual sources from this total. The Schmidt and Hunter (1977) article, as well as other articles by these authors, included computational procedures associated with the first four of the seven artifactual sources given previously. It is proposed that if the remaining variance is zero or near zero, validity generalization has been achieved, because the observed variation in validity results has been shown to be entirely a result of statistical artifacts. Further, as Schmidt and Hunter (1977) have pointed out: "in cases in which the mean of the corrected distribution is too low and/or the variance too great to allow conclusions [as to the generalizability of the validity], the corrected distribution will still be useful—as the prior distribution in a Baye-
sian study of the test’s validity [p. 530]." The procedures and results of such a Bayesian study are described in the Schmidt, Hunter, Pearlman, and Shane (1979) article.

Schmidt and Hunter, along with their colleagues, have diligently demonstrated the generalizability of results from numerous small studies covering several test–job relationships. In their initial publication (Schmidt & Hunter, 1977), they examined 114 validity coefficients relating tests of mechanical principles to performance of mechanical repairmen, 191 tests of finger dexterity related to performance of bench workers, 72 intelligence tests related to performance of general clerks, and 99 studies of spatial relations correlated with performance as machine tenders. In the Schmidt, Hunter, Pearlman, and Shane (1979) article these results were extended through the examination of generalizability of various tests related to performance in two families of clerical jobs and the job of first-line supervisor. With respect to clerical jobs, the criterion–performance relationships of 11 tests were examined, with the number of validity coefficients ranging from 53 to 321. In their most recent report (Schmidt, Hunter, & Caplan, 1981), the generalizability of validities were established for four types of cognitive tests and a weighted biographical information blank, five measures in all, in relation to performance in two petroleum industry job groups.

The results of Schmidt and Hunter’s investigations have been nothing short of a revolution with respect to validation research. In essence, they have sorted through the confusing and varying results of a 50-year period to show that a “true” validity can be established. They are of the belief that these estimates are far more meaningful than the results of typical studies with small samples for individual scientists and that validities are possible even when they are not technically feasible in the context of a particular predictor criterion relationship.

**Meta-Analysis**

The term *meta-analysis* comes from the work of Glass (1976, 1977) and involves integrating findings across studies. The idea is similar to that advanced by Schmidt and Hunter (1977), namely to bring together results from numerous small studies into an integrated study. Glass (1976) was seeking a way of organizing and depicting results from numerous studies as an alternative to the traditional narrative review. Again, the most definitive work in the area is by Hunter, Schmidt, and Jackson (1982) and describes both quantitative and qualitative procedures for integrating findings across studies. The methods are similar to those in validity generalization, namely one of removing sources of artifactual variance. However, the range of possibilities is far greater than just the simple correlation coefficients considered in the validity generalization work. Hunter et al. (1982) deal with the possibility that results of the several studies to be integrated might be presented in terms of diverse statistical procedures, such as regression, canonical correlation, or multivariate analysis of variance. In addi-
tion, procedures were considered for identifying moderator variables or interactions that are indicative of findings that might be selected to be integrated.

Meta-analysis has clearly been an innovation whose time has come. Although the original introduction of the method by Glass was 1976–1977, there have already been extensive publications using meta-analysis procedures. An extensive computer review of the validity literature for 1980 and 1981 indicated 11 and 10 articles/dissertations, respectively. This is extremely rapid diffusion, equivalent to the current diffusion of the multitrait–multimethod matrix after 20 years.

**PRODUCTIVITY ANALYSIS**

New attention has been focused on procedures that have been available for over 30 years to estimate work force productivity on the basis of validity information. Some of the original work can be traced to Brogden (1949) and the well-known publication by Cronbach and Gleser (1965), *Psychological Tests and Personnel Decisions*. More recently Schmidt, Hunter, McKenzie, and Muldrow (1979) have suggested simplified procedures that make the previously cumbersome productivity analysis approach within the range of possibilities in most situations.

The goal in productivity analysis is to estimate the dollar impact that would be realized in using a valid test to select individuals for a particular job. In the past the practical value of a selection procedure has been estimated in terms of the increase in the percentage of "successful" workers through expectancy table analysis or some equivalent procedure. Seldom have these estimates been in terms of the economic implications of the valid selection procedure on work force productivity.

The basic formula for overall gain in utility from use of a test is:

\[
\Delta U = N_s r_{xy} SD_y \bar{z}_x - \frac{N_s C}{p}
\]

where

- \(N_s\) = number of selectees
- \(C\) = cost of testing one applicant
- \(p\) = selection ratio
- \(\bar{z}_x\) = average standard score on the test of those selected (in applicant group standard score units)
- \(r_{xy}\) = test validity
- \(SD_y\) = standard deviation of job performance in dollar terms among randomly selected employees.
As shown by Schmidt and Hunter (1980), the first four items of information are easily determined. In the past it was believed that the standard deviation of job performance dollars (SD) could only be estimated using cost accounting procedures that were both complex and uncertain. Schmidt and Hunter (1980) have shown how SD could be estimated by supervisors of the job under study using a questionnaire procedure. In the Schmidt and Hunter (1980) study, budget analysis supervisors were given the following instructions:

Now, based on your experience with agency budget analysts, we would like for you to estimate the yearly value to your agency of the products and services produced by the average budget analyst. Consider the quality and quantity of output typical of the average budget analyst in the value of this output. In placing an overall dollar value on this output, it may help to consider what the costs would be of having an outside consulting firm provide these products and services [pp. 55–56].

Following an appropriate opportunity to provide that estimate, the supervisors were instructed:

We would now like you to consider the “superior” budget analyst. Let us define a “superior” performer as a budget analyst who is at the 85th percentile. That is, his performance is better than that of 85 percent of his fellow budget analysts and only 15 percent of budget analysts turn in better performances. Consider the quality and quantity of the output typical of the “superior” budget analyst. Then estimate the value of these products and services. In placing an overall dollar value in this output, it may again help to consider what the costs would be of having an outside consulting firm perform these products and services [p. 56].

Schmidt and Hunter (1980) were able to use these estimates to obtain final estimates for SD and were able to estimate the value of productivity gains from the use of a test in hiring 2000 budget analysts at over 32 million dollars.

These fairly innovative procedures for estimating the component of an important equation (SD) should make feasible the analysis of the productivity impact of selection procedures. As Schmidt and Hunter (1980) concluded: “the results of these analyses will convince many who are currently skeptical that good selection is critical to organizational success [p. 57].”

**IMPLICATIONS FOR TEST CONSTRUCTION AND TEST USAGE**

It is worth reemphasizing that validity speaks to the ultimate value of a test by affirming, or denying if that be the case, the inferential value of the score in a particular circumstance. As such, validity evidence has obvious implications for the worthiness of a test’s construction and the appropriateness of its usage.
Problems with the Process of Validation Research

Despite the importance of validity evidence, validation research has not always been of the nature that one could point to with pride. The initial half of this century could be characterized as relying most heavily on criterion-related evidence of validity, often in a way that represented “blind empiricism” at its worst. To be sure, the methods of factor analysis popularized in the 1930s encouraged the development of constructs, but the methods were somewhat prohibitive until the commercial availability of the electronic computer in the mid-1950s. Until rather recently, validity research meant a predictor–criterion correlation to the average practitioner. Even worse, as evidenced by the initial court cases on employment discrimination, tests had a half-life of their own and often enjoyed widespread use without concomitant validity evidence. In retrospect, it was clear that validation, as the feedback loop to test construction and test usage, could not operate effectively if not undertaken.

Changes in Validation Research

Change was rapid and proceeded along several fronts. The formalization of construct validation, more than anything else, legitimized validation research as a scientific undertaking rather than as a practitioner art. The definition of construct validity began in the mid-1950s and has continued in a steady, albeit slow, progression ever since. Certainly the 1979 conference discussed at length earlier showed that much progress has been made and that work continues using the diverse research strategies recommended.

There can be no doubt that Title VII of the Civil Rights Act of 1964 has been a profound stimulus in validation research. Although there was a latency period of 6 or 7 years before the Griggs v. Duke Power Co. (1971) case communicated in clear and forceful terms that tests had to measure the person for the job and not the person in the abstract, the effect has been profound.

The initial flurry of activity, at least at the practitioner level, involved efforts to validate existing tests. Implications for test usage were immediate as validation efforts failed and test programs for employee selection were discontinued. At the same time, work was initiated on alternate validation strategies. These alternatives included such diverse approaches as attempts to define and refine further content-oriented validity along with application of several of the multivariate validation models discussed previously. The obvious capstone to these many efforts has been the validity generalization research by Frank Schmidt, John Hunter, and their colleagues.

The Future of Validation Research

The future of validation research is promising. There has been more progress in the last decade than in the previous quarter century. Extending this trajectory will
undoubtedly lead to new learning about the inferential value of tests in predicting and understanding behavior.

Obviously the work on validity generalization will continue. The profession has had only a short time to adjust to these fairly unique notions. Perhaps the recency of the research is best exemplified by the fact that virtually all the work has been by Schmidt and Hunter, along with their students. Ultimately their work should render as obsolete the need for the empirical validation that has so characterized the research to date. Practitioners and researchers will only need to analyze jobs or situations of concern and, on the basis of these circumstances, consult tables of generalized validities from the numerous previous studies using various predictors in similar circumstances. This work is still in its infancy, and the implications are yet to be felt.

The inferential value of any single assessment or combination of measures is at best such to explain half the criterion variance. This is not a problem that will be addressed by the ongoing work on validity generalization or utility concepts. Instead, the multivariate validation models hold the single best hope of improving the inferential value of assessments. By seeking to incorporate information about the types of individuals and types of behaviors with organizational consideration and consequences, these models hold the best hope of improving the level of predictions. As we have seen (Owens & Schoenfeldt, 1979), these multivariate procedures can bridge the construct and empirical validity procedures. On the other hand, the procedures are complex, and progress has been slow. The hope of the future is being unshackled from the necessity of endless small studies of empirical validity with efforts being directed to the multivariate procedures.

REFERENCES


Ghiselli, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educational and Psychological Measurement*, 1960, 20, 675–684. (b)
Guion, R. M. Content validity conference. The Industrial-Organizational Psychologist, 1974, 12(1), 18. (a)
Guion, R. M. Open a new window: Validities and values in psychological measurement. American Psychologist, 1974, 29, 287–296. (b)
4. TEST VALIDATION RESEARCH


McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. The development and background of the Position Analysis Questionnaire (PAQ). Occupational Research Center, Purdue University, June 1969.


Proceedings of Content Validity II. Bowling Green, Ohio: Bowling Green State University, 1975.


Whitely, S. E. Latent trait models in the study of intelligence. *Intelligence*, 1980, 4, 97–132. (a)

Whitely, S. E. Multicomponent latent trait models for ability tests. *Psychometrika*, 1980, 45, 479–494. (b)