ROBERT M. GUION

On Trinitarian Doctrines of Validity

The evaluation of how well one measures particular attributes of people or of objects is, at least in part, a question of validity. Some discussions of validity also refer to relationships between measures of quite different attributes, either to aid in the understanding of a construct or to establish a basis for comparison between evaluations of the validity of measurement and evaluations of the validity of a hypothesis. The three conventionally listed aspects of validity—criterion-related, content, and construct—are examined from this dual perspective. The unifying nature of the validity of measurement is found in the degree to which the results of measurement (the numbers or scores) represent magnitudes of the intended attribute. Validity is thus an evaluative judgment based on a variety of considerations, including the structure of the measurement operations, the pattern of correlations with other variables, and the results of confirmatory and disconfirmatory investigations. Validity in this sense is close to the concept of construct validity but perhaps without the theoretical implications of that term; like construct validity, the evaluation cannot be expressed with a single research result. Evaluations of the validity of hypotheses should also be based on multiple considerations rather than on single coefficients. In some circumstances, conventional methods of validation may be superfluous.

People who use tests speak of "validity" in referring to evaluations either of their tests or of their use of tests; the term is ambiguous. Sometimes it refers to evaluations of how well the scores represent the attribute being measured; sometimes it refers to evaluations of how well the scores are related to some quite different attribute. Although these are complementary meanings, they are conceptually distinguishable.

Measurement that is not called testing also needs to be evaluated. Recognition of other kinds of psychological measurement may help in clarifying the concept of test validity; in return, such clarification may offer guidance for evaluating other kinds of measurement. Questions of evaluation, and certainly references to validity, are regrettably rare in reports of measurement not based on the traditional concept of psychometrics—a concept generally fenced in with the unfortunate phrase "educational and psychological testing." Examples include measures of the degrees of socialization in children's play, the information content of sentences, differences between intensities of experimental treatments, recidivism rates of mental hospital patients, the level of aggression in mice, or the degrees of preference for various classes of objects. Even in test validation studies, the list of examples includes criterion measures.

In each of these examples, something is quantified or measured, whether well or poorly, in a serious test of a hypothesis in which that "something" is important. Either the concept of validity applies to all of these different kinds of measurement or the limits of its applicability need to be better understood. Each of these examples came from a research report containing no mention of validity or any other evaluation of the ef-

0033-0175/80/1103-0385$00.75

fectiveness of measurement. Why do authors apparently ignore the question of validity?

One reason might be that the term is simply not in an author's working vocabulary. It seems likely that many authors were never exposed to a course in tests and measurement; many of those who were exposed have successfully warded off the disease.

Another reason might be disdain. Some people think mental testers are less scientific than experimenters—perhaps only a cut above astrologers. They are assumed to be holdovers from a static, discredited trait theory that is not adequate for genuinely scientific study. The attitude may be that since validity belongs to the mental testers, it can be discarded by everyone else.

Perhaps a better reason is confusion over the meaning of the word. There has been an almost mystical, trinitarian concept of validity in mental testing over the last quarter century. Although the trio of terms introduced in the "Technical Recommendations for Psychological Tests and Diagnostic Techniques" (American Psychological Association [APA] et al., 1954)—content validity, criterion-related validity, and construct validity—were identified as different "aspects" of validity, many practitioners seem to think of them as three quite different things. Such misinterpretation is "a conceptual compartmentalization of 'types' of validity . . . [that] leads to confusion and, in the face of confusion, oversimplification" (Dunnette & Borman, 1979, p. 483). This regrettable confusion and oversimplification reached its zenith with the publication of the "Uniform Guidelines on Employee Selection Procedures" (Equal Employment Opportunity Commission et al., 1978). The Guidelines seem to treat them as something of a holy trinity representing three different roads to psychometric salvation. If you cannot demonstrate one kind of validity, you have two more chances!

These three terms may have outlived their usefulness, but at the present time, no other terms serve quite the same purpose of identifying facets of validity. Clarification of the concept of validity and of its applicability beyond testing demands a clarification of the interrelatedness and the essential unity of these three terms.

The metaphor of the holy trinity is partially apt. In Christian theology, the Trinity is spoken of as one God manifested in three persons. In psychometric theology, we can speak of one validity, evidenced in three ways. The weaknesses of the metaphor are, first, that the essential unity of validity is much more closely related to the notion of construct validity than to the other two "persons" and, second, that there is reasonable doubt whether the other two consistently serve as evidence of validity.

In brief, the argument of this essay suggests:

1. Measurement consists of operations leading to quantitative statements that rep-

ROBERT M. GUION *is Professor of Industrial/Organizational Psychology at Bowling Green State University, where he was Chairman of the department from 1966 to 1971. He has written a comprehensive book on* Personnel Testing; *more than 30 articles about assessment, employment testing, personnel assessment, and measurement; and at least 7 articles on validity. He is a fellow in the Amerian Psychological Association, Divisions 5 and 14.*
COMMENTS ON AN EARLIER DRAFT *by Marvin Dunnette, Clifford (Jack) Mynatt, Patricia Smith, and Ross Stagner are gratefully acknowledged; both their comments and the differences in their perspectives were challenging.*
REQUESTS FOR REPRINTS *should be sent to Robert M. Guion, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.*

resent magnitudes of a variable conceptualized by the researcher.

2. The validity of the measurement consists fundamentally of the congruence of the operational and the conceptual definitions of the variable.

3. Evidence of such congruence may be partially based on relationships of the variable, as measured, to other variables. The evaluation of the strength of such relationships may be sought, however, even where the quality of the measures is not questioned. In principle, the validity of a hypothesized relationship can be studied independently of studies of the validity of measurement.

4. For some kinds of measurement, the traditional questions of validity and of so-called validation strategy do not arise. For these, different questions may be formed to evaluate the measurement and its use.

The Validity of Measurement

One does not measure objects or people; one measures attributes of objects or people. Much of everyday physical measurement, such as height, is based on mathematically formal procedures of fundamental measurement (Coombs, Dawes, & Tversky, 1970; Torgerson, 1958). There is general agreement among scientists and others on the definition of the standard unit of measurement. The evaluation of how well one has measured is based on a judgment of the precision of measurement (by which is meant either the fineness of the permissible error or the reliability of repeated measurement) and of the accuracy of measurement (by which is meant a freedom from biasing tendencies to err systematically too high or too low). In these cases, validity (as typically defined) is not an issue.

This is not to say that validity is irrelevant to physical measurement. Much of the work done by physicists, astronomers, and others in the "hard" sciences uses measurements of one kind of physical phenomenon as a basis for inferences about a different one. For example, Idso, Jackson, and Reginato (1975) described methods of using the ratio of reflected to incoming solar radiation as a basis for inferring soil moisture by remote surveillance from high altitude aircraft or satellites. Validity of such inferences is indeed an issue and has been addressed by Idso et al. (1975).

It is an issue much like that in the psychological measurement of "softer" attributes, such as aspects of intelligence or personality, where there are no units of measurement with standard definitions accepted by the scientific community. Although we speak of reliability (another ambiguous term), the concept of validity is invoked as the definitive basis for evaluating such measurement.

In short, whether the attribute being measured is physical or psychological, "hard" or "soft," the focus of measurement is necessarily on the attributes—the "something" that is measured. The something may be clearly identified or not, well established in the scientific literature or innovative, important or trivial, material or abstract; it can be a vague idea, a moderately well-defined concept, or an established scientific construct. Nothing in the intent to measure it says that the attribute must be clearly defined at the outset. To the contrary, many scientific constructs develop in an iterative pattern of rough definition, preliminary measurement, and refinement of definition. The point is that one needs at least a hazy conceptual definition of the something to be measured.

One then develops a set of operations, an operational definition, by which relative

magnitudes of the defined something can be assessed. A set of operations must first be evaluated logically in terms of its apparent relevance to the underlying concept. If it is deemed satisfactory in this respect, it can be put to use and, after use, evaluated. Validity in measurement is, first of all, essentially an evaluation of how well one has succeeded in measuring the attribute that was to be measured.

The operational definition in measurement ends up with a number—a scale value, a score, or some other expression of quantity. The number ordinarily has no absolute meaning in its own right. It becomes important when it is used to draw inferences about the attribute being measured. Validity, then, refers to an evaluation of the quality of the inferences drawn from these numbers.

It follows that the validity of measurement is not a precise evaluation. Rather, it is expressed in broad quantitative categories: high validity, satisfactory validity, or poor or no validity. One might compare validities of inferences based on different operational definitions of an attribute and say that the validity of one is better, or equal to, or worse than the validity of inferences from another. These are ordinal statements; they do not denote precise quantities, and they are not expressible with precise numbers. *One should not confuse an evaluative interpretation of validity with an obtained validity coefficient.* Validity coefficients may be computed, but the evaluation of validity is based on those coefficients and on other information as well; it is not equated with them.

Validity is a property of inferences from scores, not (strictly speaking) of the measuring instrument or test itself. Properties of a test should influence evaluations of validity, but so also should other information. If one is evaluating a total approach to measurement, evaluative considerations include structural characteristics of stimulus materials, the degree of standardization, the adequacy of the sample taken in measurement, and the like. These things contribute to one's evaluation of the validity of certain inferences from scores, but they should not be confused with those inferences.

CRITERION-RELATED VALIDITY

The many reasons for conducting studies relating one set of measures to another can be condensed into two categories: (a) to investigate the meaning of scores as measures of a certain attribute and (b) to investigate the scores as concomitants or predictors of *other* attributes (APA et al., 1974).

The first of these fits the old definition of validity as the extent to which a test measures what it "purports" to measure. If one has developed a test "purporting" to measure scholastic aptitude, then the "real" measure of that aptitude is achievement in school (Hull, 1928). Pintner (1931) advocated validation of intelligence tests against such other indicators of intelligence as teachers' ratings, school achievement, or even scores on other intelligence tests. These were the criteria—the standards—for judging the goodness of the test as a basis for inferences about intelligence; for example, the test is evaluated favorably as an intelligence test if the correlation between test scores and school achievement is high, and it is evaluated less favorably if that correlation is low.

The second category is illustrated when one correlates school achievement with scores on an intelligence test that one has *already* evaluated as providing satisfactorily valid measures of intelligence. The purpose might be a practical interest in predicting

achievement; intelligence may be one of several attributes investigated as potential predictors. In this kind of investigation, the so-called criterion is placed less in the role of a standard and more in a role like the dependent variable in an experiment. The analogy is useful because, in such criterion-related validities, the inference from the test score is based on a hypothesis. The hypothesis is that performance on the test is related to performance on the other measure, usually a measure of a totally different attribute and usually one of greater importance to the test user. In these cases, validation is not so much a matter of evaluating the test score for measuring some attribute as an evaluation of a hypothesized relationship of one variable to another.

In personnel testing, the hypothesis is that an attribute of job applicants, as measured, can be used to predict future proficiency (or attendance, or whatever) if the applicants are hired. The future proficiency is of greater interest to the organization than the attribute that predicts it; the "validation" research is an attempt to evaluate the hypothesis that proficiency is a function of the predictor. The study may, of course, lead to insights about the measurement of the trait. If, over a period of time, several studies are conducted in which the same sort of dependent variable is predicted by scores on the same test, then there is a pretty good basis for nailing down one's interpretation of the meaning of scores on that test. In a pragmatic world, however, that is often treated as a relatively trivial bit of information. The interpretation of interest to the organization is, quite simply, the value of the test as a basis for predicting future performance and, therefore, as a basis for decisions, regardless of what attribute it really measures.

The two different purposes of criterion-related validity studies can be distinguished in another way. If one's purpose is to come to a better understanding of how well a particular attribute is being measured by a certain test, then research should consist of several studies, each using a different criterion thought to reflect that attribute. If one is primarily interested, however, in predicting a specific measure of future behavior, then the advisable strategy is to use many different predictors; for example, measures of proficiency may be hypothesized to be functions of certain applicant attributes, situational variables, and demographic variables used in combination (Guion, 1976).

CONSTRUCT VALIDITY

The first type of criterion-related validation yields evidence of what has been called construct validity. To evaluate the measurements for inferences of specified attributes, some criterion measures are chosen as independent reflections of the same attributes. Other criterion measures may represent competing interpretations of what a particular test score (or other measure) might mean, that is, different attributes.

For example, suppose that one considers scrap rate (proportion of work that is done so poorly it must be discarded—scrapped) an indication of poor work motivation. That is, it is proposed that a worker's scrap rate can be used as a measure of carelessness; perhaps the idea is that careless people should not be promoted to a higher level of responsibility. If one is seriously interested in evaluating the notion that scrap rate measures carelessness, it would be well to look for correlations with some other indicators of carelessness, but it would also be well to check relationships with some measures that indicate clumsiness. Clumsiness is a competing interpretation of scrap rate; the scrap rate cannot be considered a valid measure of carelessness if scrap is

largely attributable to lack of coordination. This may still be important if the higher level job is another job requiring high levels of coordination. However, if motivation (in the sense of attentiveness) rather than physical grace is required on the next higher level job, the use of scrap rate would probably be an invalid instrument for selection in part, at least, because it is an invalid measure of the desired attribute.

The essential logic of construct validation is disconfirmatory. There should, of course, be positive evidence that a measurement procedure leads to valid inferences about a particular construct, but the issue is more commonly a matter of showing that alternative or competing inferences do not destroy the intended interpretation. Cook and Campbell (1976) described construct validity primarily in terms of freedom from experimental confounding. In the correlational language of psychometric discussions, it is perhaps more familiar to say that a construct has been measured validly if a set of scores is reasonably free from contaminating sources of variance. The aim of research in construct validation is to strengthen, if possible, a given interpretation of scores, assuring that alternative interpretations are not very good. Of course, if the alternative interpretation turns out to be a very good one, the originally intended interpretation may have to be modified.

The historical introduction of the notion of construct validity was as much concerned with the validation of a theoretical construct as with the validation of its measure (Cronbach & Meehl, 1955). The basic logic and disconfirmatory emphasis of construct validation, however, can be as useful in evaluating attributes and measures of attributes identified vaguely for purely practical purposes as for evaluating constructs and measures of constructs required in the development of a theory. Perhaps, all that is being implied here is a metaphor, an analogy to the original notion of construct validity. If so, the analogy is apt.

To say that valid inferences can be drawn about a specified construct by a particular method of measurement is to say very little about the value of that measurement for practical decisions. In personnel selection, the practical value of measurement depends not on how well it measures a specified attribute but on how well it predicts future performance on some other variable. Evidence of that practical value *may* come from a criterion-related validity coefficient of the second kind described in the previous section. In the long run, better evidence may come from a tightly reasoned hypothesis coupled with strong evidence of the construct validity with which the independent variables of that hypothesis are measured (Guion, 1976).

Issues of the validity of measurement arise in basic experimental research as well as in testing. The measurement problem arises whenever a concept is imperfectly or partially operationalized, and it becomes an acute problem whenever an experiment fails to confirm a theoretical proposition. In such cases, the experimenter must ask whether the failure is because the axiomatic relationships posed by the theory are wrong or because the inferences drawn from the measurements are invalid. Stagner (Note 1) has given me an excellent illustration. He and Harlow did the first curare experiment and concluded that animals could not learn a striped muscle response when paralyzed. Later studies showed that they had learned but that the learning could be shown only when curare was injected again. The data were not in error, but the inference was.

CONTENT VALIDITY

Content validity is also a special case of construct validity. The "construct" may be

an attribute, like level of knowledge or level of skill, in a particular information or performance domain. It has been customary to speak of content validity when one wishes to use scores on a test to infer probable performance in a larger domain of which the test is but a sample.

In personnel testing, the concept of content validity, which was borrowed from educational measurement, has been very troublesome. In educational measurement, a test could be considered a valid measure of curriculum content insofar as the material covered on an examination matched in general proportions the material to be covered in the general curriculum. In either case, the so-called content validity of the test is an evaluation of how well the tasks or questions it contains match those in a defined content domain. In personnel testing, the definition of a content domain has been a source of very great confusion. Nowhere is that confusion better documented than in the "Standards for Educational and Psychological Tests" (APA et al., 1974). In discussing the applicability of content validity to employment testing, it says that "the performance domain would need definition in terms of the objectives of measurement, restricted perhaps only to critical, most frequent, or prerequisite work behaviors" (p. 29). Two paragraphs further, it says, "An employer cannot justify an employment test on grounds of content validity if he cannot demonstrate that the content universe includes all, or nearly all, important parts of the job."

A Strict Approach to Content Sampling

The procrustean task of making content validity fit the problem of personnel testing can be described by a four-step process that would assure a work sample test of unquestionable job relevance:

1. Define a *job content universe* on the basis of job analysis. This should include all nontrivial tasks, responsibilities, prerequisite knowledge and skill, and organizational relationships that make up the job. This is *not* what is to be sampled directly. One rarely hires people who are already able to do all of the things that are done on the job. (The second of the two quotations given earlier is herewith declared unacceptable.) Training programs exist to teach people how to recognize and carry out job responsibilities. Job applicants may be expected to know already how to do some of the things the job requires. For example, in hiring a secretary, one expects to train the new employee in specific office procedures or the use of unique equipment encountered in that office, but one does not expect to teach the new secretary how to type.

2. Identify a portion of the job content universe for the purposes of work sample testing; this may be called *job content domain*. The word *domain* is being used here to denote a sample—not necessarily a representative sample—of the content implied by the word *universe*.

3. Define a test content universe as the tasks to be included in testing and the possible methods to standardize and score performance on them. The test content is not merely a sample of job content; it includes things that are not part of the actual job. Performing a job and taking a test are not the same thing, even if the component tasks seem nearly identical. Typing mailable letters from dictation on a real job involves a familiar machine, knowledge of the idiosyncrasies of the person who dictates the letters, telephone or other interruptions, and so forth. Typing the same material in a test situation involves the anxiety or motivation created by the testing, standard conditions such that distractions (if any) are built into the exercise equally for all people taking the test,

and using material dictated by an unfamiliar voice. Moreover, typing on the job is not formally scored, but the test requires a standard scoring procedure. Therefore, one adds to the job content domain possible methods of standardization and of scoring to form the test content universe. It consists of all the tasks that might be assigned from the job content domain, the various conditions that might be imposed, the various procedures for observing and recording responses, and the possible procedures for scoring them. The test will not include everything, but defining such a universe identifies the options.

4. Define a *test content domain,* a part (not a representative part) of the test content universe. This defines actual specifications settled on for test construction. A test constructed according to these specifications would certainly be seen as job related. If the measurement operations were not questioned, it might even be said to have a high level of content validity.

The foregoing steps define a very tiresome and exhaustive procedure, but they should make clear two points: (a) that what has been talked about as content validity is really a content-oriented approach to test construction (Messick, 1975) and (b) that a truly representative sample of the job does not ordinarily provide measurement of the quality of performance.

A test constructed by this procedure will almost certainly result in valid inferences about the ability to do the job, but the evidence of that validity may require something beyond the implications of the term *content validity*. In the first place, it is highly unlikely that the two domains would precisely overlap; if circles are drawn to represent each, they would overlap, but the degree of content validity (the degree of overlap) would be small if either the job tasks omitted or the measurement procedures added were substantial. In the second place, there is an important conceptual difference between evaluations of the validity of inferences from scores and the evaluations of the quality of sampling tasks. Content validity, by definition, refers to the latter.

If the inference to be drawn from a score on a content sample is to be an inference about performance on an actual job, then it is drawn at the end of a series of inferential steps, any one of which can be a serious misstep. The most serious misstep may occur in defining the scoring system. The scoring system of a work sample is as subject to contamination as is the scoring of any other test. The score obtained by an individual may reflect the attribute one wishes to infer—ability to do the designated aspects of the job; but it may also reflect a variety of contaminations such as anxiety, ability to comprehend verbal instructions, or the perceptual skills that enable some people to perceive cues for scoring that are imperceptible to others. All of this has a familiar ring after the discussion of construct validity. It means that disconfirmatory research (that is, construct validation) may be needed to evaluate the validity of scores on many job samples. To repeat: Content validity is a special case of construct validity (Messick, 1975; Tenopyr, 1977).

A Unitarian Doctrine of Validity

This discussion, of course, has been purely semantic, offered in the conviction that semantic clarification leads to clearer thought. The meaning of validity begins with a concept of an attribute, more or less clearly defined. It has ended with an evaluation of how well such a concept is represented by the numerical result of a set of operations for measuring it. Content validity, insofar as its meaning is restricted to content sam-

pling, may influence one's evaluation of the validity of inferences from numbers or scores, but it is conceptually distinguishable from this broader concept of validity. Criterion-related validity sometimes gives evidence directly relevant to this representation question, but sometimes its evidence is more directly relevant to evaluations of the tenability of a hypothesis—again, a conceptually distinguishable idea. Stated differently, both the kinds of evidence known as content validity and as criterion-related validity may contribute to evaluations of how well the operations represent the underlying concept, but they do so only insofar as they are special cases of construct validity. Construct validity seems to provide the unifying theme.

I am a little reluctant, however, to assert that validity in general is the same thing as construct validity. Discussions of construct validity have generally been carried on at a level of the philosophy of science, and not all evaluation of measurement needs to be done at this level. At the risk of hedging, therefore, I identify the unifying concept of validity as similar, but not necessarily identical, to what has been meant by construct validity.

Validity is therefore defined as the degree to which the result of the measurement process (the numbers) satisfactorily represent the various magnitudes of the intended attribute. This is familiar; it is another statement of the traditional definition of validity as the extent to which a test measures what it purports to measure. It is not, however, restricted to tests; the emphasis is on a more general evaluation of the goodness of measurement.

That evaluation can draw from all three aspects of validity. Certainly, what has been called content-oriented test construction contributes to the evaluation of the adequacy of measurement. If the results of measurement are to be called valid, structural questions about the measurement operations must be answered satisfactorily. These include, but are not limited to, questions of content. Questions of content apply not only to work samples but to factored aptitude tests, rating scales, personality inventories, and just about any other technique. These are not questions of the representativeness of the content and measurement operations, however; they are questions of importance and relevance. In discussing the validity of criterion measures, Jenkins said,

> There is always the danger that the investigator may accept some convenient measure . . . only to find ultimately that the performance which produces this measure is merely a part, and perhaps an unimportant part, of the total field performance desired by the sponsor. (Jenkins, 1946, pp. 96–97)

To generalize: There is always the danger that the measure at hand is so constructed that it cannot faithfully mirror the attribute to be measured.

In addition to content, structural considerations include reliability, standardization, language and language level, quality of graphics, and appropriateness of time limits or of standardizing samples, among others. In short, a first approximation to a judgment that a particular set of operations leads to valid inferences about a specified attribute is the judgment that the set of operations has been thoughtfully and skillfully assembled. This may not be a sufficient basis for a judgment that the measures are valid, but it is a necessary one.

Some form of empirical evidence is equally necessary. The evidence typically takes the form of a pattern of correlations. The measures being evaluated should logically be correlated with some external variables, but there are others to which they should logically *not* be related. The judgment of validity depends on the confidence one has

that obtained coefficients fit the logically expected pattern. Individually, such correlation coefficients are statements of criterion-related validity; collectively, they are bases for judgments of construct validity.

It seems clear that the essence of a unified conception of validity is at least metaphorically the notion of construct validity; in short, the trinitarian doctrine reduces to a unitarian one so long as the meaning of validity refers only to the evaluation of how well a designated attribute is measured.

The Validity of a Hypothesis

The discussion of construct and criterion-related validation includes a different kind of evaluative research, the evaluation of hypotheses about relationships of either theoretical or practical importance. Since such research is often called validation, even if the purpose is not to validate the measurement of an attribute, it is useful to speak of evaluating the validity of a hypothesis. This requires evaluation of the research as well as the result. Under some circumstances, the validity coefficient obtained in the research is inflated. For example, there may be common error variance in both the predictor and the variable to be predicted. Under most circumstances, however, problems in the research lead one to underevaluate the validities of one's hypotheses.

Campbell and Stanley (1966) and Cook and Campbell (1976) discussed the validity problems encountered in doing experimental and quasi-experimental research; much of their discussion is also relevant to nonexperimental studies of relationships. In addition to construct validity, they referred to "internal validity" in discussing problems in the conduct of research that undermine permissible confidence in the results. They referred to "statistical conclusion validity" in discussing statistical issues that alter such confidence. They referred to "external validity" in describing problems limiting the generalizability of research findings. In these discussions, Campbell and his associates identified what might be called a third target of evaluation, the validity of the research itself. Such evaluation is surely necessary in establishing the validity of a hypothesis.

Validation research in personnel testing is usually correlational rather than quasi-experimental. This seems unfortunate. We should have broad programs of personnel selection that can eventually be evaluated for overall effectiveness, without such severe concentration on the single predictor applied to the individual applicant. Use of program evaluation designs could adopt such dependent variables as organizational productivity or profitability—variables not predictable when the individual is the unit of analysis.

We need not wait, however, for the adoption of program evaluation methods to consider the effects of the internal, external, and statistical conclusion validities on the evaluation of correlational results. These considerations should make clear that the particular validity coefficient one obtains in a predictive study is not a sufficient basis for an evaluation of the validity of the hypothesized relationship (any more than it is sufficient for evaluating the validity of the measurement). Many of the threats to validity described by Campbell and his colleagues conspire in prediction research to lull the researcher into either an unwarranted complacency or an unwarranted pessimism. If their effect lowers the estimate of the relationship, the researcher avoids using a good predictor.

Personnel testers have placed too much reliance on criterion-related validation and on the specific validity coefficients they obtain for evaluating their predictive hypotheses. The simplicity of the validity coefficient makes it very attractive; predictive studies, where they can be done well, are obviously valuable sources of data. However, things are rarely as simple as they seem, and it is time to abandon a simplistic overreliance on a validity coefficient obtained from a single study. There are several reasons.

First, research conditions are never exactly repeated. The logic of research on the validity of a predictive hypothesis generally assumes a static set of conditions such that the particular setting in which the study is done this year will be matched in all but trivial respects by the setting in which the results of the study will be used 2 or 3 years hence. The typical design of research and blind use of the results seem to assume that new methods of training, new equipment, new social attitudes, new characteristics of the applicant population, and many other new things will have no effect on the observed relationship.

Second, the logic assumes one or more variables important enough to predict. Too often validity studies use available criteria without serious evaluation of their importance. Jenkins (1946) called for criteria that were comprehensive measures of the performance of concern to the "sponsor" of the research; he decried the still-prevalent tendency to accept any measure that happened to be lying around. Otis (1971) and Wallace (1965) argued that psychologists should develop behavioral criteria instead of the typical managerial records or ratings. The advice is not often taken.

Third, the logic assumes that the measures to be predicted will be psychometrically sound, that is, that they will represent consistent behaviors reliably observed and that they can be measured validly. However, it is very rare that the report of a criterion-related validity study says anything at all about the evaluation of the criterion measures themselves. The use of supervisory ratings is prevalent, and the validities of these ratings are often questionable.

Fourth, the logic assumes that the relationship observed will generalize to later samples. If motivation or attitude influences predictor scores, as in personality inventories or measures of physical strength, the findings in a concurrent study (using present employees with assurances that poor performance will not haunt them) may not generalize to samples of job applicants.

Finally, the logic assumes reliable statistics. Criterion-related studies should be conducted, if at all possible, using reliable measures encompassing a representative range of talent on many more cases than are usually available (Schmidt, Hunter, & Urry, 1976).

In short, the evaluation of a research result must always take into account the adequacy of the sample (its representativeness and size), procedures and safeguards in research (e.g., avoidance of criterion contamination), the logical and psychometric quality of the measures of the variables, and the rational foundation for the hypothesis (Guion, 1976). Certainly, one should take into account the relevant history of prior research.

The latter point suggests a less static approach to prediction research. At a recent convention, Croll and Urry (Note 2) described a Bayesian approach in which each new sample provides new information to be assimilated in the light of prior information about probabilities; an address on the use of Bayesian statistics in industrial psychology was also presented by Novick (Note 3). The Bayesian approach seems an effective way to point out that a single validity coefficient is not as useful for evaluating the

tenability of the hypothesis of a predictor–criterion relationship as is a series of such coefficients (Schmidt & Hunter, 1977).

Must All Tests Be Validated?

The heading is a paraphrasing of the question asked by Ebel (1961); his answer was negative. I think he was right. This does not mean that measurements and hypotheses should not be evaluated; it means that there are other methods and standards for evaluation beyond those implied by the conventional trinitarian doctrine of validity.

The unifying concept of the validity of measurement has been defined in terms of the congruence of the conceptual and operational definitions of an attribute; more precisely, validity is the degree to which the numbers obtained by a measurement procedure represent magnitudes of the attribute to be measured. Fundamentally, like the notion of construct validity, this definition refers to the meaningfulness or interpretability of the scores. By this definition, all measures (including tests) should be valid; it does not follow that all measures should be validated by looking at content sampling, or validity coefficients, or experimental or multivariate studies of convergence.

Consider, for example, the most fundamental sorts of measurement, such as the measurement of weight using balances or the measurement of linear distance with a yardstick; consider also such mathematically formal measurement as the measurement of information or uncertainty. For these kinds of measurement, there is a formal mathematical model representing the finite set of relationships involving the attribute. Effective formal measurement will provide an isomorphic measurement set, that is, a measurement set with a one-to-one relationship to empirical realities or to the theoretical model. Such isomorphism may be sufficient evidence of the meaningfulness of the measurement that psychometric concepts of validation are superfluous. (For further discussion of the evaluation of formal measurement without reference to psychometric validity, see Coombs, Dawes, & Tversky, 1970; Hooke, 1963; or Suppes & Zinnes, 1963.)

It has also been suggested that the validity of a hypothesized relationship between variables be accepted as conceptually distinguishable from the validity of measurement. Here, too, there are circumstances in which validation research is not necessary and perhaps not meaningful. For example, a content sample does not need to be forced into the notion of either kind of validity to be considered job relevant. Under certain conditions, operational definitions of an attribute, such as ability to do a job, provide both a necessary and a sufficient evaluation of the obtained scores and of their use in personnel selection without further concern for either kind of validity.

Particularly in personnel research, the procrustean concept that everything must somehow be squeezed into a validity framework needs to be questioned. I have heard colleagues seriously propose, for example, that educational, or experience, or even age requirements be defended on the grounds of content validity! The principal concern in personnel testing—as in most fields of applied psychology—is with the validity of a hypothesis of a relationship between a variable used as a predictor and subsequent job performance. If solid research to evaluate the hypothesis can be conducted, complete with valid measures of job performance, then the research conventionally called criterion-related validation provides the best evidence of the usefulness of the measure—its job relatedness.

It would be an error, however, to assume that job relatedness can be evaluated only in terms of a validity coefficient describing an observed relationship. The validity coefficient itself must be logically evaluated. Beyond that, the solid logic of a well-developed hypothesis, where competent empirical research is unlikely, provides better evidence of the job relatedness of a predictor than does a validity coefficient obtained in a faulty study.

Validation is important, but it is not all-important. Sound arguments of the job relevance of well-measured, logically defensible attributes may be sufficient in themselves.

REFERENCE NOTES

1. Stagner, R. Personal communication, February 23, 1979.
2. Croll, P. R., & Urry, V. W. Tailored testing: Maximizing validity and utility for job selection. In T. Kehle (Chair), *Innovations in personnel selection.* Symposium presented at the meeting of the American Psychological Association, Toronto, Canada, August–September 1978.
3. Novick, M. R. *Implications of Bayesian statistics for industrial/organizational psychology.* Invited address presented at the meeting of the American Psychological Association, Toronto, Canada, August–September 1978.

REFERENCES

American Psychological Association, American Educational Research Association, & National Council of Measurement Used in Education (joint committee). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin,* 1954, *51,* 201–238.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. *Standards for educational and psychological tests.* Washington, D.C. American Psychological Association, 1974.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental design for research.* Chicago: Rand McNally, 1966.

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally, 1976.

Coombs, C. H., Dawes, R. M., & Tversky, A. *Mathematical psychology: An elementary introduction.* Englewood Cliffs, N. J.: Prentice-Hall, 1970.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, *52,* 281–302.

Dunnette, M. D., & Borman, W. C. Personnel selection and classification systems. *Annual Review of Psychology,* 1979, *30,* 477–525.

Ebel, R. L. Must all tests be valid? *American Psychologist,* 1961, *16,* 640–647.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. Uniform guidelines on employee selection procedures. *Federal Register,* August 25, 1978, *43*(166), 38290–38315.

Guion, R. M. Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally, 1976.

Hooke, R. Some figures. In R. Fox, M. Garbuny, and R. Hooke (Eds.), *The science of science.* New York: Walker, 1963.

Hull, C. L. *Aptitude testing.* Yonkers, N.Y.: World Book, 1928.

Idso, S. B., Jackson, R. D., & Reginato, R. J. Detection of soil moisture by remote surveillance. *American Scientist,* 1975, *63,* 549-557.

Jenkins, J. G. Validity for what? *Journal of Consulting Psychology,* 1946, *10,* 93-98.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist,* 1975, *30,* 955-966.

Otis, J. L. Whose criterion? In W. W. Ronan & E. P. Prien (Eds.), *Perspectives on the measurement of human performance.* New York: Appleton-Century-Crofts, 1971.

Pintner, R. *Intelligence testing: Methods and results* (2nd ed.). New York: Holt, 1931.

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology,* 1977, *62,* 529-540.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validity studies. *Journal of Applied Psychology,* 1976, *61,* 473-485.

Suppes, P., & Zinnes, J. L. Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1). New York: Wiley, 1963.

Tenopyr, M. L. Content-construct confusion. *Personnel Psychology,* 1977, *30,* 47-54.

Torgerson, W. S. *Theory and methods of scaling.* New York: Wiley, 1958.

Wallace, S. R. Criteria for what? *American Psychologist,* 1965, *20,* 411-417.

Received April 16, 1979