LEVEL III (12)

ARI TECHNICAL REPORT
TR-79-A9

Principles of Work Sample Testing:
II. Evaluation of Personnel Testing Programs

by

Robert M. Guion

BOWLING GREEN STATE UNIVERSITY
Bowling Green, Ohio 43403

AD A072447

April 1979

Contract DAHC 19-77-C-0007

Prepared for

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

79 08 06 088

U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

WILLIAM L. HAUSER
Colonel, US Army
Commander

NOTICES

DISTRIBUTION Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333

FINAL DISPOSITION This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report TR-79-A9	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): PRINCIPLES OF WORK SAMPLE TESTING. II. EVALUATION OF PERSONNEL TESTING PROGRAMS.	5. TYPE OF REPORT & PERIOD COVERED: Final rept. 15 Nov 76 - 15 June 78
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Robert M. Guion	8. CONTRACT OR GRANT NUMBER(s): DAHC19-77-C-0007
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Bowling Green State University, Bowling Green, Ohio 43403	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q161102B74F
11. CONTROLLING OFFICE NAME AND ADDRESS: US Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, Virgina 22333	12. REPORT DATE: April 1979
	13. NUMBER OF PAGES: 62
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): --	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

Monitored by G. Gary Boycan, Engagement Simulation Technical Area, Army Research Institute.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Measurement theory, psychometrics, work sample testing, validity, content-referenced testing, criterion-referenced testing, latent trait theory, generalizability theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Principles are offered for increasing the objectivity of measurement in programs of personnel testing. Classical concepts of reliability and validity are reviewed. Construct validity is seen as the basic evaluation of a measuring instrument in psychology; criterion-related validity actually refers to hypotheses rather than to measurements, and content validity refers to test development. The major evaluation for personnel tests is less a matter of validity than of job relevance and of generalizability. Implications of latent trait theory and generalizability theory are discussed in terms of content-referenced

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

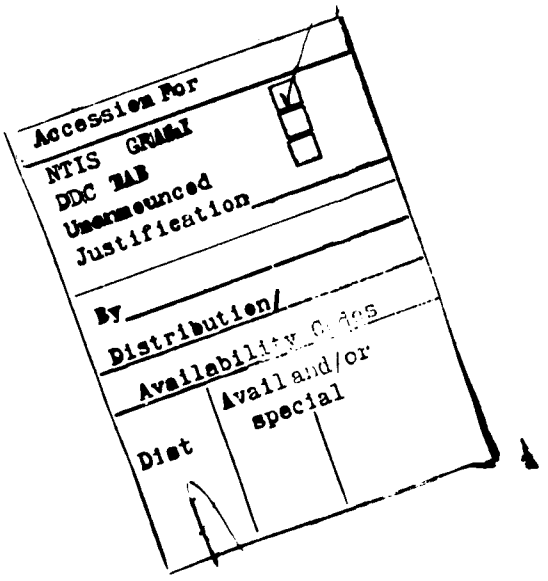
UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (continued)

testing for work samples.

This report, the second of four, is written for psychologists interested in testing and psychometrics in general.



PRINCIPLES OF WORK SAMPLE TESTING: II. EVALUATION OF PERSONNEL TESTING PROGRAMS

BRIEF

Personnel testing should be as objective as possible. Objectivity in measurement occurs under two conditions: if the scale does not depend on who has been measured with it, and if the measures do not depend on the specific scale used. If the stimulus-response content of the test permits verifiable responses, if the format imposes no constraints on the responses, and if the responses are free from distortion, the principle of objectivity is approached.

There are aspects of the test, however, other than its stimulus-response content. Scoring procedures which are defined without respect to the content may be attached to it; inferences are often drawn going far beyond the content. The more objectively attributes can be measured, the less reaching is needed to make appropriate inferences and, therefore, the less elaborate the research needed to evaluate the measurement.

Classical concepts of reliability and validity are reviewed. Criterion-related validity is noted as concerned with inferences about other variables rather than inferences about the measure used as a predictor; criterion-related validity therefore does not evaluate the measurement per se, although it evaluates hypotheses about predictor-criterion relationships. Construct validity is seen as the essence of validity, and it is defined in terms of the proportion of total variance explainable by the construct being measured. The essence of construct validity research is disconfirmatory; that is, it is intended to consider alternative interpretations of the meaning of scores which, if supported, would disconfirm the originally proposed inferences.

Content validity is not really validity at all; it is an evaluation of the procedures of test construction, not of inferences drawn from scores. In personnel testing, the test development procedure can lead logically from definitions of a job content universe and domain to the definition of a relevant test content domain and establishment of test specifications; if the test is constructed according to those specifications, its job relevance is virtually assured. Under certain circumstances, the assurance of job relevance is all that is needed in evaluating a personnel test.

Alternatives to classical psychometric theory are examined for potential value in personnel testing, especially in work sample testing. Work sample tests are seen as being, by definition, content-referenced

tests. Latent trait theory is examined for its implications for scaling personnel tests, and the implications of generalizability research are also considered.

It is concluded that too much attention is given to classical concepts of validity and not enough to the more immediately important evaluations of job relatedness and of generalizability beyond the testing situation.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

The preceding paper in this series surveyed psychological measurement in general. It concluded with the idea that different kinds of measurement of different kinds of variables, and perhaps for different purposes, demand a different emphasis in evaluation. This paper will also be quite general, although with explicit references to the central problem of work sample testing, as it describes important considerations in the evaluation of personnel testing programs. This discussion assumes that personnel testing is best understood as taking place in settings of institutional control, even if actually done as field research, and that it covers the gamut of variables and methods of measurement.

GENERAL CONSIDERATIONS IN THE EVALUATION OF TESTING PROGRAMS

The emphasis is on the total evaluation of a total program; the evaluation of a testing program includes but should not be limited to validation. In some circumstances, conventional questions of validity may not arise at all; where valid inferences from scores must be ascertained, positive research results may be sufficient, but negative results leave many unanswered questions. A total testing program consists of offering a test under a standard circumstance as a stimulus, obtaining and scoring responses, and drawing inferences from the scores for the sake of making personnel decisions. Only the latter directly uses classical validation procedures.

THE TEST AS STIMULUS

The test content, instructions, administrative procedures, format, and the situation in which the test is administered all contribute to a stimulus complex which should be standardized; evaluation of a testing program should inquire first into the details of its standardization. Are instructions given according to clearly standard procedures? If so,

are they uniformly understood by all examinees before the test begins? If there are time limits or other constraints on performance, are they rigidly standardized and enforced, as they should be?

The principle basic to these and virtually all other questions in evaluation is straightforward: does a person's score on the test represent clearly the attribute being measured, or do other attributes of the person, the test, the procedure, or the setting in which the testing is done have some influence on the score? To the extent that irrelevant attributes influence obtained scores, a testing program is in some sense deficient.

The second line of inquiry concerns the degree to which the content, format, structure, and technique contribute to the objectivity of measurement. The term objectivity has been indiscriminantly applied in psychological measurement without much precision of meaning; when people refer to objective tests, they frequently mean multiple-choice tests. It is true that such tests may be objectively scored, but the measure becomes less than objective to the extent that the available options either constrain or suggest the responses of the examinee.

The most objective measurement is mathematically formal and is best illustrated by physical measurements. Two characteristics of such measurement make it genuinely objective: (a) the scale exists independently of the objects used in developing it, and (b) the measurement is independent of the particular instrument used for measuring. In presenting these requirements for objectivity, Wright said by way of illustration, "But when a man says he is five feet eleven inches tall, do we ask to see his yardstick?" (Wright, 1968, p. 87).

By these standards, traditional psychological testing is never objective. The measures (scores) depend on the particular sets of questions asked and on the sample of people (objects) used in item analysis, and

the meaning depends on the sample used in establishing norms. By analogy, however, some characteristics of objective measurement can be approximated in even traditional psychological testing. In objective measurement of the length of objects, for example, the measure can be verified, the instrument imposes no constraints on the results of measurement (save in the fineness of calibration), and the object itself cannot distort the measurement. By analogy, since psychological measurement is based on responses, a test is objective to the extent that its content permits responses that can be verified, and it places no constraints on the nature of the responses, and that the responses are undistorted. The reference to the multiple-choice format as "objective" suggests a further analogy in that the observer who reads the yardstick or who scores the test should not be able to distort the results.

The first contribution of the test as stimulus to objectivity is its content. A test of arithmetic skill problems can be far more objective than a measure calling for endorsements of statements of belief, partly because it is less ambiguous. In any content area, the objectivity of measurement can be enhanced if the content domain to be sampled is clearly defined and the procedures for sampling clearly specified. Clarity in defining the domain and the procedures for sampling is insurance against ambiguity. Ambiguity of content may distort responses, and it will surely lead to unreliable inferences from them.

In evaluating the test as stimulus, no special constraints need to be placed on the nature of the domain to be defined. It may be a performance domain, a domain of factual information, or a domain of approaches to measurement. For example, one may set out to construct a test of problem-solving ability. Literally dozens of problem-solving tasks may be used. One might use block designs, small assembly tasks, manipulative tasks, exercises in logical reasoning, and countless others. The domain of problem-solving is not a very unified domain. If one wanted to sample for the problem-solving test all possible kinds of problem-solving tasks,

the result would be not only an incredibly long test but one that would lack internal consistency -- an important evaluative consideration. The domain, therefore, might be defined specifically in terms of the use of anagrams. Within this domain, boundaries can be established and the characteristic tasks available for sampling can be identified. One might specify that the domain will consist of seven-letter anagrams of from two to five vowels. Other specifications can be added. With the domain clearly defined, a test constructor can establish rules for sampling the domain. The rules may specify only procedures for sampling the content domain, or they may specify statistical rules for accepting items sampled. For example, if the test is to be used for conventional norm-referenced interpretations, it may be specified that item difficulties are to be within a given range and that all item-total correlations must be above some minimum value. The result of clear specifications of the test domain should be clearer meaning of scores.

These points should not be over-emphasized in an evaluation. An excellent testing program may be based on serendipitous findings using haphazardly constructed tests. Nevertheless, the final overall evaluation of the testing program is more likely to be favorable if the stimulus properties of the program have been carefully constructed.

THE TEST AS RESPONSE

The content of a test, and therefore the content of the domain sampled, is a stimulus-response content. The test is not the printed instructions or questions or assigned tasks; it is the combination of instructions and success in following them, questions and answers, or tasks and performance. Objectivity of traditional measurement from responses to stimuli, already discussed, is illustrated in Figure 1.

If the response options for the examinee are wholly open-ended, the testing content is defined in part by the content analysis of the responses

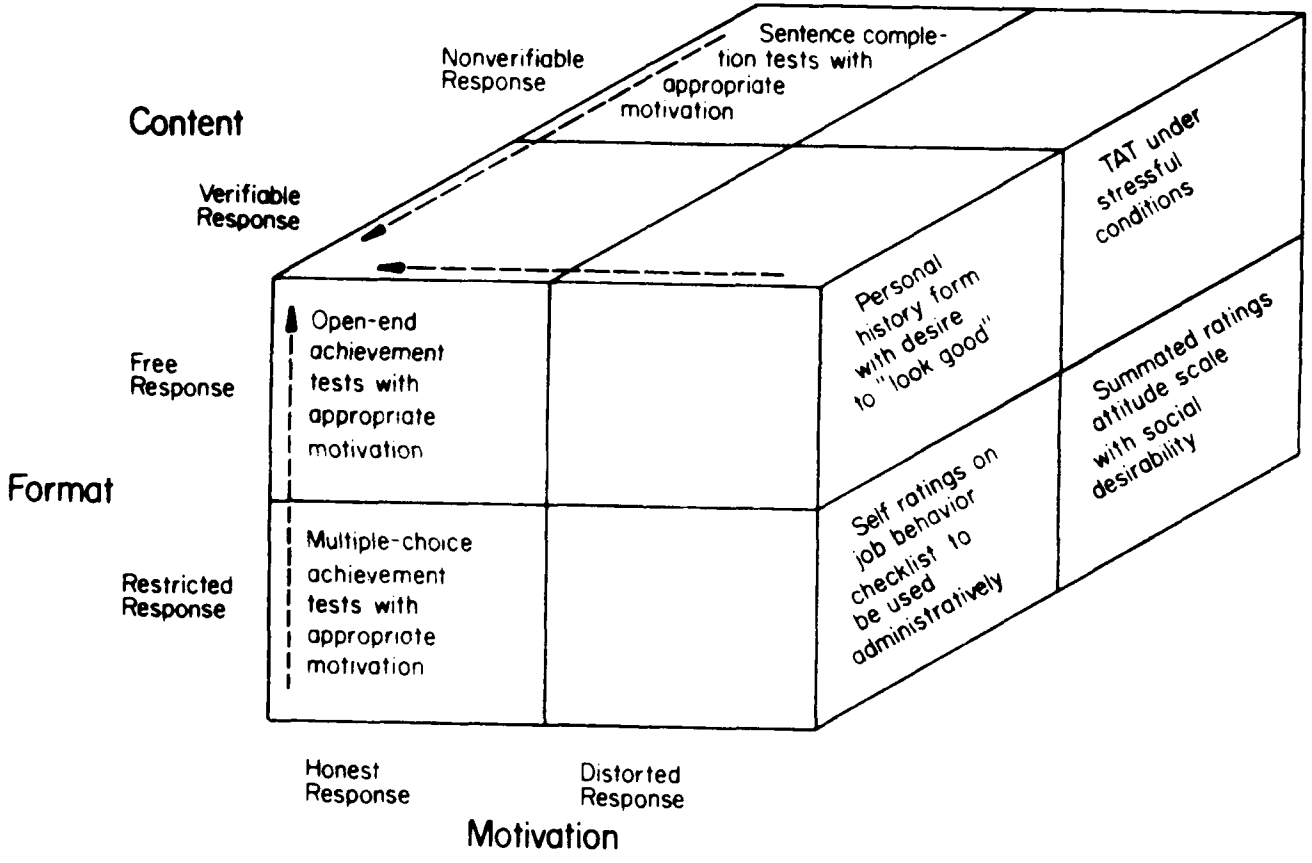


Figure 1. Schematic diagram showing contribution to objectivity of three response dimensions (adapted from Guion, 1965)

and the resulting scoring categories. Most performance tests involve open-ended responses. A work sample test, for example, consists of telling the examinee to do something. The actual responses (that is, the performance) may be observed and classified, or the consequences of the response (that is, the product of the behavior) may be evaluated along selected dimensions. The measurement is, however, objective in format, if not in scoring, because of the unrestricted opportunities for response. Objectivity may suffer (because of the necessity to classify responses), however, if the procedure permits observer or scorer characteristics to influence an obtained score intended to be a measure of an attribute of the examinee. Maximum objectivity in measurement may require an optimal tradeoff between the distortion created by artificially restricting responses and the distortion created by the unreliability or bias of observers' classification and scoring procedures.

If the form of the test is in a restricted response mode, such as multiple-choice, the definition of the test domain should include possible or plausible responses. It is a useful practice, and an indication of great care, to begin the construction of a multiple-choice test by administering the items in open-ended form to a substantial sample of people. The responses used to complete an item stem can be tallied, and a domain of potential responses can be identified with rules for selecting the correct and distracting options.

Responses to test items or tasks must yield scores if there is to be any measurement. This is one of those underwhelming, obvious kinds of statements that often seems to be overlooked. The point is that the stimulus-response content does not often include the score. In a multiple-choice test, the traditional scoring is simply a count of the number of items answered correctly, but this is a traditional convenience, not dictated by the content domain. In other forms of examination, such as work samples, the scoring procedure may have to be invented. In either case, the scoring procedure needs to be evaluated both for classical

reliability and for the possibility of contamination in the scores.

THE TEST AS INFERENCE

This heading includes the traditional concern for validity in the evaluation of testing. The topic will be examined in more detail in a subsequent section; it is sufficient here to indicate the possible variety of inferences and the evaluative questions of validity they pose.

One form of inference, according to the APA Standards (APA, AERA, and NCME, 1974), is the inference of performance in a domain based on performance on the sample. Evaluation of this sort of inference has been based on the ambiguous notion of content validity.

A different type of inference involves inferring performance on one measure from performance on a different one. Evaluation of such inferences is based on criterion-related validity.

In the third class of inferences, an individual's standing on some underlying characteristic presumably measured by the test is inferred from the score. Evaluation of such inferences is based on construct validity.

Different inferences are sought for different purposes, and therefore the emphasis on evaluating the validity of inferences is likely to differ in different testing situations. Nevertheless, it should be understood that virtually all mental measurement involves to some degree all three kinds of inference. On the anagrams test of problem-solving discussed earlier, to evaluate the inferences as valid, the evaluator must be willing to infer that performance on that set of anagrams is a good indicator of performance on any other set of anagrams from the same specified domain, he must be willing to infer that performance on the anagrams task is related to performance on something else of particular interest to the evaluator, and he must be able to infer that the performance on the

anagrams fits a network of relationships in which the scores can be interpreted in terms of problem-solving ability rather than in terms of some other characteristic such as verbal comprehension.

THE TEST AS A TOOL FOR DECISION

Most personnel testing is done to provide a basis for decisions, not primarily to measure an attribute. Evaluation of the test as a measuring instrument is important to its evaluation as a decision tool, but the two evaluations should not be confused. An excellent measure may be a poor basis for decision; a poor measure may nevertheless be the best decision tool available.

Decisions are based on predictions, either literal or implied. In personnel testing, therefore, the usual and primary evaluation of a testing program lies in the magnitude of the correlation between scores on the test and subsequent measures of the variable to be predicted. Even in situations where it is either infeasible or unnecessary to compute such a correlation coefficient, the logic of trying to maximize an implied predictive relationship remains the paramount basis for evaluating decision tools.

A principal implication of that logic is the general rule that complex performance can be predicted better with a set of predictors than with any one test. In most practical personnel prediction problems, a test battery will be devised, and some form of composite score will be computed for each person. In all discussions of test scores that follow, this composite score is as relevant as a score on a single test.

Multivariate prediction does not necessarily or uniformly imply a composite. The different variables might be arranged in some sort of sequence of decisions. Where this procedure is followed, one evaluates the testing program, and any particular test or test composite within it,

in the light of its position within the sequence. Its position in the sequence becomes another aspect of the setting in which the test is given.

The decision to be made is not an automatic consequence of the prediction. A cutting score may be set, above which individuals are selected or certified (or whatever), and it may fluctuate from time to time according to changing standards or to supply and demand. Subjective considerations may influence decisions independently of test scores. In some settings, variables that may have influenced obtained test scores may be considered by applying a mathematic correction or some sort of subjective fudge factor. Whether the decision is based solely on test scores, or whether other considerations influence the decision, the decision itself is the final step in the testing process to be evaluated. According to modern decision theory, the evaluation should be based on concepts of utility and cost effectiveness. A comparison of the costs of Type I and Type II errors should be made in evaluating the utility of the decisions.

Although the logic of prediction is almost always implied in any personnel decision, the arithmetic of prediction may be superfluous. The logical prediction frequently made, particularly when testing for a particular skill, is that a high-scoring person will perform better by using that skill if placed in a job or a training program that demands it. For example, it is almost an unarguable proposition that a person who scores high on a test of typing skill will be able to handle the typing assignments of the ordinary office. There is neither any need to compute a criterion-related correlation coefficient, nor is there much desire for it, since the extraneous factors that might inhibit performance (such as immediate conflict with the supervisor) are of little interest in the evaluation of the testing program. In these kinds of situations, the test score is interpreted on its own terms. One who types more words per minute is assumed to be able to type more words per minute than someone else. The score is its own operational definition of a skill that is prerequisite to successful performance on a job.

If the conditions of performance on the job are substantially different from the conditions of performance in the testing situation, a question of generalizability arises. To use an absurd but descriptive example, individual differences in a standardized typing test might have very little relationship to individual differences in performing the same typing task in a pitching rowboat.

The example is an extreme example of the problem of the generalizability of test scores. A permissible inference under one set of conditions may not be permissible under a quite different set of conditions. A major evaluation for many decisions is whether the generalizability of performance in the test situation to performance in the targeted conditions is a reasonable assumption. If there are to be dramatic differences in conditions, then empirical verification of generalizability yields important information.

The issue of fairness, which has been central in most discussions of personnel testing since the passage of the 1964 Civil Rights Act, should be understood as a special case of generalizability. In the situation where tests are evaluated with criterion-related correlation coefficients, the issue is one of the generalizability of the regression equation; do the constants computed for a composite of all groups apply equally well to any identifiable subgroups? In tests which are evaluated without such correlations coefficients, it may be more important to identify and evaluate the magnitude of various sources of error. Is a measure of performance on a work sample, for example, influenced by an observer's knowledge of the race or sex of the examinee? If so, to what extent? Is the task so organized that persons of unusual height have a handicap in the test situation that would not influence performance under more realistic conditions; that is, has the standardization of the test created an artificiality that influences the scores of some people unfairly because it does not exist in other conditions?

CLASSICAL PSYCHOMETRIC THEORY: RELIABILITY

It has been said that all measurement, regardless of method or attribute measured, must be reliable. It does not follow, however, that the identical ways of examining reliability apply in all cases.

The essence of an investigation of reliability is an assessment of the degree to which variance in a set of measurements may be attributed to error. Different kinds of variables, and different methods of measurement, are susceptible to different sources of error. Moreover, different methods of estimating reliability are sensitive to different sources of error (Stanley, 1971).

With many forms of measurement, one is especially interested in the stability of the description over time, that is, in errors due to instability of measurement. However, error over time is not relevant to all measurements. Blood pressure, for example, is not stable over time; it varies according to activity level, tension, etc. Yet failure to find the same blood pressure under these different conditions would never be considered to be an error of measurement; it is simply a valid reflection of the changes that occur in the attribute being measured. On the other hand, measures that are supposed to represent relatively enduring traits, such as behavioral habits or personality characteristics, should stay rather constant, at least in reasonably similar conditions, over some substantial period of time. Variations over time in measurement in these cases constitutes error. Changes in obtained measurement over time are likely to be considered sources of error for measures of personality traits, cognitive skills, motor skills, job knowledge, and most measures of performance or proficiency. It is probably inappropriate to treat change over time as a source of error in measurement for most physical or attitudinal variables.

A general principle in measurement is that one should measure one attribute at a time. The standard way of measuring, exemplified by typical tests, is to use many fallible operations to measure the same thing and accumulate observations. Thus a test will consist of many items, each with different specific content but each presumably tapping or reflecting the same fundamental attribute, -- that is, each a miniature test. The total test is a sample of observations from a homogeneous universe. Behavioral statements in rating scales constitute a similar example, as, perhaps, do pieces of information obtained through records. Obviously, observation of behavior over time can likewise be divided into "items." In short, tests can often be said to consist of component parts, each of which constitutes an independent observation of the same variable. If, however, one of these components proves to reflect an attribute other than that being measured, the inclusion of that component in the total leads to an error of measurement. An item that measures something different from the rest of the items is a contaminating item. An observation taken during a time period where a sharp noise or other distraction occurs is a contaminating observation since it reflects behavior under distraction rather than behavior under attention to the task at hand. It is conventional to refer to studies of errors in sampling the observations as estimates of internal consistency or homogeneity in measurement.

Homogeneity should be a pervasive concern in all measurement, and it is virtually assured in fundamental measurement. Tests, on the other hand, represent very small samples of nearly infinite populations of possible items and are especially susceptible to such sampling errors. To investigate these errors, it is common practice to develop and compare parallel forms. As a matter of fact, classical reliability theory assumes parallel test forms meeting rather stringent definitions. Because it is unlikely that non-test approaches to measurement will meet the requirements for parallel forms, domain sampling error in these methods may be substantially larger.

An analogous reliability problem occurs when substantially different (i.e., non-parallel) methods are used to measure the same attribute. It probably matters little whether one measures the length of a board with a flexible steel tape measure or a wooden yardstick, but in some areas of physical measurement the alternative approaches are anything but parallel. Measuring distances in cartographic analysis by triangulation is in no sense parallel to measuring with a ruler. If the two methods give different results, one or both of them may be wrong, a simple correlation to demonstrate that the methods are inconsistent is not a sufficient basis for assigning error to either one.

If observers are the instruments of measurement, there is probably a finite number (greater than one or two) of possible observers. The one or two actual observers used are samples from the universe of possible observers and, as such, may be sources of measurement error.

If two observers are used, each may contribute a unique error or measurement, and it is necessary to determine the degree to which any composite measure is subject to error in sampling observers and the degree of such error should be assessed. There is no way to estimate the error due to sampling observers if only one observer is used, just as there is no way to ascertain error attributable to a specific set of questions if only one form of the test is used. Repeated samples of observers (or tests) are required to estimate the degree to which the sampling introduces error into the measurement. Likewise, if ratings are used, some error may be due to the raters chosen, and it can only be evaluated by determining the degree of agreement between raters. Another example calls for estimating agreement among scorers of open-ended test items.

Many forms of measurement involve a subjective assignment of people or objects to scaled categories. An attempt to measure aggressive tendencies under conditions of provocation in an assessment center, for

example, mi ht present an assessee with an anger-producing situation, and observers may be instructed to determine whether the response fits better in a category described as turning white and silent, or a category described as verbal expressions of anger, or a category described as using physical movements symbolic of attack. Such measurement poses two quite different kinds of reliability problems. One is the degree to which observers may agree on their observation of behavior; the other is the degree to which the numbers assigned to the categories fall along a reproducible scale. (The point of view taken here is that the Guttman index of reproducibility is a special case of reliability.)

Still another potential source of error can be broadly identified as a condition of measurement. The results of measurement may be different if the measurement is taken in the morning or in the late evening, it may be different if it is taken under sanitary, optimal conditions rather than in less pleasant but more realistic field conditions. Knowledge of the extent of such errors may often be useful, even if not often available.

Gross estimates of most of these sources of error can be estimated by conventional methods of estimating reliability by internal consistency coefficients, coefficients of equivalence, coefficients of stability, or coefficients of agreement (conspect reliability). However, these coefficients simply do not do a particularly clean job of separating out the components of error associated with attributes of the person, method, or setting other than the attribute measured. If one computes an internal consistency coefficient, a stability coefficient, and an equivalence coefficient, determines the proportion of variance attributable to each of the three kinds of error implied by those coefficients, and adds them up, the total estimate of error variance obtained is far greater than is realistic. A preferred approach is the generalizability analysis, or multiple facet analysis, advocated by Cronbach, Gleser, Nanda, and

Rajaratnam (1972). Using analysis of variance designs, such analysis can examine the components of error most likely to be problems in a specified testing program.

Generalizability studies seem especially useful for estimating errors in evaluating performance on a work sample. Such performance may be a function of the instructions given, the person who administers or scores the test, the activities preceding testing, and the environmental setting in which performance is being measured. One may evaluate these sources of potential error, with explicit estimates of the proportionate total variance attributable to each source, through the use of analysis of variance designs.

CLASSICAL PSYCHOMETRIC THEORY: VALIDITY

Validity refers to an evaluation of the quality of inferences drawn from test scores, qualitative summaries, judgments, or other measurements. The first point of importance in that statement is that validity is not a fact; it is an evaluation. Moreover, it is a quantitative evaluation. It is best to think of validity as expressible only in broad categories: high validity, satisfactory validity, or poor or no validity. Depending on the context, one may compare validities and say that validity in one circumstance is better, or equal to, or worse than validity in another. Since such statements do not denote precise quantities, they are not expressible with precise numbers. One should not confuse an evaluative interpretation of validity with an obtained validity coefficient. Validity is not measured; it is inferred. Although validity coefficients may be computed, the inference of validity is based on such coefficients, not equated with them.

There has been a kind of colloquial shorthand in psychometric English in which people tend to speak of the "validity of a test." Informed

people do not mean that phrase literally. It is simply a shorthand phrase referring to the evaluation of the inferences one draws from scores obtained on a test. (In this paper as in others, the shorter phrase will undoubtedly be used frequently, but there should be no misunderstanding about its meaning.) Speaking precisely, validity refers to evaluations of specific inferences that may be drawn from scores, not to evaluations of properties of tests, and there are as many validities as there are inferences to be drawn from the scores. In evaluating the total testing program, many test properties should be evaluated, such as degree of standardization, adequacy of content sampling, and the like. These properties may contribute to one's evaluation of the validity of certain inferences from scores, but they should not be confused with such inferences.

Validation refers to the processes of investigation from which the validity of certain inferences from scores may itself be inferred or evaluated. All validation procedures are in some sense empirical. Some of these procedures involve correlating test scores with other data, comparing correlations, doing experimental studies to determine differences in scores for groups differing in attributes or treatments, or evidences of procedures used in the construction of a test. Where the evidence of validity is drawn from correlations of scores with other measures, the validation does not consist simply of computing the correlation coefficient; it consists of the entire research process, including sampling of persons and of situations, the evaluations of other forms of validities of the measures used, the evaluation of the logic of the hypothesized relationship between the variables, and the purely procedural care with which data were collected. These are also empirical events, and the argument for interpreting scores on the one test as permitting valid inferences about the variable measured by the other one is supported (if at all) by the entire chain of empirical evidence, not just the correlation coefficient.

To say that certain kinds of inferences from scores are valid inferences, therefore, implies not only the empirical process of gathering data but the logical process of evaluating all of the avai' ble evidence.

Implied in the foregoing is a final observation about the nature of validity: in classical psychometric theory, validity refers to a set of scores. The evidence upon which validity may be claimed applies to the score of a single individual only if that score can be interpreted with reference to an entire set of scores. That is, in classical interpretation of scores, the individual score is considered more or less valid only if it has been previously determined that a set of scores from other individuals tested in the same way is a more or less valid set of scores. Validity is therefore defined in terms of variances; validity is the proportion of total variance relevant to the purposes of testing; irrelevant sources of variance reduce validity. A correlation coefficient describing the relationship of one measure to another is simply a means of describing the shared variance.

In short, to make judgments about the validity of the inferences one may draw from a set of scores is to make judgments about the irrelevant components in a set of scores. Earlier discussions referred to evaluations of single scores as the degree to which a score is free from reflections of attributes other than the one intended. The classical way to ascertain that freedom is to determine the level of irrelevant sources of variance. This discussion of validity in general, therefore, has reflected, without explicitly referring to them, the aspects of validity identified in the Standards for Educational and Psychological Tests (APA, et al., 1974).

CRITERION-RELATED VALIDATION

At the most directly empirical level are the criterion-related validities, predictive and concurrent. For convenience, the many reasons

for conducting criterion-related validity studies can be set in two categories: (a) to investigate the meaning that may be attached to scores on a test, that is, to identify more clearly the variable or variables measured, and (b) to investigate the utility of the scores as indicators or predictors of other variables.

The first of these grows out of the historical definition of validity as the extent to which a test measures what it "purports" the measure. If one has developed a test "purporting" to measure scholastic aptitude, then the "real" measure of aptitude is how well one does in school (Hull, 1928). School performance is then the criterion of how good the test is. That is, the correlation between scores on the test and grades in school is an index of the success of the test in measuring what it was supposed to measure. The same logic is sometimes found in modern instances in which a test of, let us say, verbal ability is correlated against supervisory ratings of verbal ability.

This kind of validation, although it involves computing a correlation between scores on the test being validated and another measure called a criterion, is better discussed under the heading of construct validity. That is, in the more conventional language of the last quarter century, such criterion-related studies are done for the purpose of verifying the interpretation of scores in terms of designated constructs.

It is an obvious outgrowth of concern for criterion-related validity that one finds that the criterion of "real" aptitude is often a variable of great importance, and the utility of the test as a predictor of that criterion becomes a matter of greater interest than the theoretical interpretation of the scores themselves. Common practice uses the term criterion-related validity primarily for those situations where one wishes to infer from a test score an individual's standing on some variable of interest that is different from the variable measured by the test. The latter variable has been called a

criterion for historical reasons, but it is usually better described as a variable analogous to the independent variable in experimental studies. The analogy is useful because, in criterion-related validities, the inference is based upon a hypothesis. That is, on a priori grounds, the test user or test developer hypothesizes that performance on the test is related to performance on some other measure, often of a different variable. Validation in such cases is less a matter of checking an intrinsic interpretation of test scores than of conducting research on the hypothesis.

In the field of personnel testing, at least for selection, the hypothesis takes the form that scores on the test can be used as indicators of potential proficiency, or some other performance variable, on a job. For example, on a given production job where each spoiled piece represents a monetary loss to the employer, scrap rate is a fundamental measure of an economic variable. With some validation, one might draw inferences about psychological variables from scrap rate (clumsiness or carelessness are competing interpretations), but this is usually not the salient point. The point is that each spoiled piece costs the organization money. If it can be shown that a particular dexterity test, or perhaps a particular test of knowledge, can predict individual scrap rates within reasonable limits of error, then the scores on the tests may be used to "infer" (more accurately, to predict) scrap rates on the job, even though the individual has not yet been trained or placed on the job. The fact that a theoretician can find an explanation for the common variance in the two sets of measurements is relatively trivial in most cases; rarely is there any attempt to interpret such criterion-related validity coefficients theoretically: what is interpreted is the value of the test as a basis for predictions of future performance. What is commonly called test validation is, therefore, best understood as an investigation of a hypothesis rather than an investigation of variables underlying scores on either predictor or criterion.

It is useful to distinguish between hypotheses that imply predictive

validity and those for which concurrent validity is appropriate. To illustrate the difference, consider the possible finding that measures of self-confidence are substantially and significantly correlated with proficiency ratings of leadership. Three independently testable hypotheses are possible: (a) that people who are self-confident become effective leaders, (b) that people who are effective leaders become self-confident, and (c) that people who are effective leaders are self-confident. The first two of these are predictive hypotheses; they predict in opposite directions. Ignoring the possibility of reciprocal causality, both of these hypotheses require predictive studies to validate them, but the design of the studies would be substantially different. In the first hypothesis, one would administer the measure of self-confidence prior to people gaining experience in leadership roles. For the second hypothesis, one would not obtain the measure of self-confidence until people have been in the leadership role long enough to establish clear and observable habits of leadership. For the third hypothesis, the two measures could be taken concurrently. The fact that very little benefit may accrue to anyone from such concurrent correlation is beside the point; the point is that the hypothesis is a different one and that in any correlational study relating them, the procedures of investigation will be different.

There has been an over-reliance on criterion-related validation in the history of personnel testing. The simplicity of the validity statement makes it very attractive, and it is often necessary for specific personnel purposes. However, things are rarely as simple as they seem, and many factors make over-reliance on a single, obtained validity coefficient questionable.

First, the conditions of a validation study are never exactly repeated. This is especially evident in the case of a predictive study, where the logic of predictive validation assumes that the conditions at the start of the study will be reasonably well matched by the conditions at the start of a new time sequence when the results of the original study are

to be applied. If a validation study extends over three or four years or more, new methods of training, new equipment, new social attitudes, new applicant characteristics, and many other new things may change the validity before the results can be put to use.

Second, the logic of criterion-related validity assumes a valid criterion. Very rarely, however, do criterion-related validity reports give any evidence of the validities of inferences drawn from the criterion measures themselves. All too often, personnel testing uses unvalidated supervisory ratings as the criterion. In many of these cases, a criterion-related validation study is probably inadvisable.

Third, the logic of criterion-related validity assumes that the sample of an applicant population used for research is truly representative and that the validity will generalize to later samples. This is almost always violated to some degree, if only through bias in attrition. Statistical procedures can, of course, provide better estimates of population validities than those provided by the biased sample, but the assumptions for these procedures often are not satisfied.

Finally, results of criterion-related validity studies, particularly those in which the predictor is a composite of several variables, are highly questionable if based on small numbers of cases. The sample size necessary to conduct a competent investigation of criterion-related validity is much larger than was earlier supposed (Schmidt, Hunter, & Urry, 1976).

CONSTRUCT VALIDITY

Despite the foregoing warnings, studies of criterion-related validity are basic in investigations of construct validity. Where the criterion is chosen because it can shed light on the intrinsic meaning of the scores being validated, such studies enable one to sharpen possible interpretations

of test scores and to choose between competing interpretations. In this context, low validities can be as helpful as high validities if they indicate what the test does not measure and thereby limit the nature of the variables legitimately inferred from the scores.

Construct validity is not a utilitarian notion. It is studied because one wishes to increase his understanding of the psychological qualities being measured by the use of a particular test. Such studies influence the degree of confidence one may have in the accuracy of descriptive inferences about the individual tested. A test is ordinarily supposed to be a measure of something; that something is an idea or concept of a variable; if sufficiently sophisticated scientifically, it is called a hypothetical construct. The latter term is intended to emphasize an idea that has been constructed as a way of organizing knowledge and experience -- that is, a construct is a work of scientific imagination. As evidence accumulates about a construct, the idea may change.

The essential logic of construct validation is disconfirmatory. One does research designed to disconfirm an intended interpretation by persistently trying alternative interpretations; that is, one investigates the possibility that a variable other than the one intended to be measured (other than what a test "purports" to measure) is a better interpretation of the scores. Variance in a test intended to be used for inferring problem-solving abilities may in fact be substantially contaminated by variance due to individual differences in reading ability. Or a newly proposed construct may prove to be an old variable conventionally measured by other means. In either case, the aim of the research is to strengthen, if possible, a given interpretation of the test by showing that alternative interpretations are not feasible. Of course, if the alternative interpretation turns out to be a fairly solid one, then perhaps the originally intended interpretation is the one that is infeasible.

The notion of a hypothetical construct in its usual context is a fairly sophisticated scientific construct itself. Reference in discussions of hypothetical constructs deal with "nomological networks" of scientific lawfulness (Cronbach & Meehl, 1955). The logic and disconfirmatory emphasis of construct validation can, however, be very useful for ideas that are much less well developed scientifically. Supervisory ratings of work proficiency can, for example, be evaluated in terms of construct validity. In this case, the construct is not a highly developed creation of scientific imagination; it is a rather vague idea of proficiency on a specific job. The question is not the scientific import or sophistication of the idea, but whether proficiency is a reasonable interpretation of the variable measured by the ratings. Disconfirmatory research would consider alternative explanations. Perhaps the ratings merely measure how long ratees have been known; therefore, studies would be initiated to determine the relationship of length of acquantance to the ratings. This is, of course, a complex question. A mere correlation between length of acquaintance and ratings may identify bias, or it may show that experience does in fact count on that job. These are competing inferences from a correlation, and the logic of construct validity requires that one attempt to evaluate them and to choose between them. In some circumstances, this might require another research study. In other circumstances, it may merely require an exercise in logic; if the job can be learned in a few days, or if proficiency is limited by external forces such as supply of material to the worker or the speed of a conveyor belt, the hypothesis that greater experience results in greater proficiency is probably silly and the correlation would disconfirm the desired interpretation of the ratings.

To say that valid inferences can be drawn about a specified construct is to say little or nothing about the utility of the measure for practical decisions. In a personnel selection situation, for example, the practical utility of the measure depends less on how well it measures a given construct

than on how well the scores will predict future performance, regardless of what or how many constructs they reflect.

As pointed out in the preceding section, in many circumstances criterion-related validity is not feasible. In these situations one exercises the logic rather than the arithmetic of predictive validity. Elsewhere (Guion, 1976), the author has used the term "the rational foundation for predictive validity" for situations in which construct validity is evaluated as part of that logic. The phrase implies that the logic of construct validation and the logic of predictive validation meet if a predictive hypothesis is very carefully developed. The steps of careful development include careful job analysis, rational inferences from the information obtained in the job analysis about the kinds of constructs that may be hypothesized as relevant to performance on the job as it would, if it could, be measured, and finally the identification of predictor variables that will validly measure those constructs. Such a logical argument pools a great deal of empirical information: the observations of the job, the group judgments involved in inferring the constructs, and the evidence of the construct validities of the predictors. None of this empirical information is necessarily expressed as validity coefficients, yet to infer that high scores on the predictors predict high performance on the job is arguably more valid under these circumstances than when a validity coefficient is obtained from an inadequate study.

CONTENT VALIDITY

Content validity is a special case of construct validity. It is likely to be emphasized in measuring knowledge or performance variables, and it is especially frequently invoked in evaluations of work samples. For that reason, it will be considered here in particular detail.

The "construct" when one speaks of content validity is more obvious

than where reference is to more abstract constructs: level of knowledge, level of skill, level of competence, or degree of mastery of a specified content or skill domain. It has been customary to speak of content validity when one wishes to infer from scores on a test reflecting the probable performance in a larger domain of which the test is a sample. We have already referred to domain sampling in sampling the kinds of items that measure a construct; the concern here is for domain sampling where the domain is more intuitively understood.

Content validity began in educational measurement as a straightforward concept which posed no special problems. An educational curriculum identifies an explicit body of knowledge and instructional objectives, and educational practice has decreed that asking a question about specific knowledge is an acceptable operation for measuring it. Therefore, if one had all possible questions about a specified curriculum content, one could obtain a universe or domain score by adding up the number of items answered correctly. When one takes a sample of all possible items from that domain, one can add up the number of items answered correctly and, from that score, infer something about the number or proportion of items that would have been answered correctly had the entire domain been used.

This account is perhaps unnecessarily glib, but the glibness gives it brevity. It is acknowledged that the best practice in sampling content domains defined by educational curricula would utilize what Cronbach (1971) called the universe of admissible operations, which identifies stimulus-response content in terms of the permissible kinds of questions and the expected kinds of responses. Nevertheless, the glibness, if that is what it is, seems defensible because the universe of admissible operations in educational testing is reasonably restricted. A combination of curriculum identification and conventional practice relieves many questions that might otherwise arise.

In personnel testing, however, the concept of content validity has been much more troublesome. The definition of a content domain has been a source of great confusion, and it is therefore necessarily difficult to define a universe of admissible operations for measuring a domain one does not clearly understand. Perhaps nowhere is the confusion better documented than in the Standards (APA, et al., 1974). In its discussion of the applicability of content validity to employment testing, that document points out 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." Two paragraphs further, on the same page, we read, "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" (p. 29).

Job Content Universe. In attempting to clarify matters, it may be useful to distinguish between the terms universe and domain and between job content and test content. We may, therefore, identify four conceptual entities: a job content universe, a job content domain, a test content universe, and a test content domain.

A comprehensive job analysis may identify all the nontrivial tasks, responsibilities, prerequisite knowledge and skill, and organizational relationships inherent in a given job, and all of this defines a job content universe.

Tasks are the things people do; job analysis need not identify trivial tasks, but it should identify the most salient activities. Responsibilities may include tasks but may also include less clearly observable activities. A teacher, for example, may be responsible for the health and safety of the children in her class. The precise activities carried out in fulfillment of that responsibility may be hard to define since they vary with changed circumstances. Prerequisite knowledge

and skill represent cognitive or motor abilities or information necessary for effective and responsible task performance. Such knowledge or skill needs to be defined unambiguously; vague trait names are not enough. "Must be able to compute means, standard deviations, correlation coefficients, and probability estimates" is a far more explicit statement than saying, "Must have knowledge of statistics."

Organizational relationships place the job in its context; they identify systems of ideas, materials, or social relationships as they influence the job; dependencies that may exist in sequences or task performance, and the degree to which people in other jobs must depend on the incumbent in the job being analyzed in doing their job. Both the organizational relationships and the responsibilities describe not only the content of the job but the content of the consequences of the performance of a job.

If the job is at all complex, it would be either impossible or absurdly impractical to try to develop a work sample test to match that total job content universe, it might be necessary to carry out the full training, to provide experience, and to observe performance on the actual job for a period of time. If one's purpose were selection, this would be absurdly impractical.

Job Content Domain. In practice, one identifies a portion of the job content universe for the purposes of testing. In a stenographic job, for example, the portion of the universe most salient in selection or performance evaluation might be restricted to those aspects involving typing. From the job analysis, one could identify the tasks, responsibilities, and the prerequisite skill (such as spelling) associated with typing; with these restricted elements, and ignoring other aspects of the total job content universe, one can define a task content domain. In this sense, the word domain is being used as a sample (and not necessarily a representative one) of the content implied by the word universe.

Test Content Universe. Performing a job and taking a test are not identical activities, even if the component elements are identical. To continue with the stenographic example, typing mailable letters from dictation on a real job involves interruptions, knowledge of the idiosyncracies of the person who has dictated the letters, interruptions by telephone calls or requests for materials from files, etc. Typing from the same dictated material in a test situation involves typing under the anxiety created by the testing and its peculiar motivational characteristics, in standard conditions such that any distractions are built into the exercise and are standardized for all people, and using material dictated by an unfamiliar voice. To the best of this writer's knowledge, no one has ever developed a typing test that is a genuine work sample in the sense of duplicating actual circumstances, distractions, and snide comments on the dictation tape -- nor has he encountered anyone who would advocate it.

Instead, one defines from the job content domain a universe of possible operations for the development of a test. The test content universe, therefore, consists of all of the tasks that might be assigned, all of the conditions that might be imposed, and all of the procedures for observing and recording responses that might be used in the development of the content sample. The test content universe is, again, a sample of the job content domain. But it is more than that; it includes elements that are not part of the job content domain since the latter probably includes no information about procedures for observing and recording behavior on assigned tasks. This would be particularly true if the operations decided upon consisted of a series of questions about the reasons for certain procedures in carrying out a task; one would virtually never include such question-and-answer exercises as a part of the actual job, but they can be quite useful in testing people to determine their qualifications for the job.

This is a crucial point in the total chain of argument. In many kinds

of work sample testing, psychometric considerations require the inclusion of non-job components in defining a test content domain; otherwise, there may be no measurement of anything. Such added operations may involve ratings by observers, counting (and perhaps weighting) responses to questions, or carrying out physical measurements and inspection of products that go beyond those encountered in the actual job itself but are necessary foundations for measurement.

Test Content Domain. The test content domain is a sample of a test content universe, and it defines the actual specifications for test construction. Again, the test content domain is not necessarily a representative sample of the test content universe. Questions of practicality and of relative importance must assuredly enter into the judgments defining a test content domain.

There remains, then, the actual construction of the test.

The Limits of Content Sampling as Validity. The foregoing sequence, which is illustrated by Figure 2, is not necessary as a detailed procedure, but the four-step process of domain definition is useful for clarifying the relationships of job and test domains and for reconciling the contradictory statements in the Standards.

It should be clear that what has been called content validity is quite different from all other forms of validity. As a matter of fact, the term should not be used since it can only cause confusion. The term validity refers, as has been pointed out, to an evaluation of the inferences that can be made from scores. If the inference to be drawn from a score on a content sample is to be an inference about performance on an actual job, it is drawn at the end of inferential leaps, in any one of which there can be a serious misstep. The crucial chance for misstep is in the definition of a test content universe; it is here that a system of scoring (or its basis) is invented, and that system of scoring is

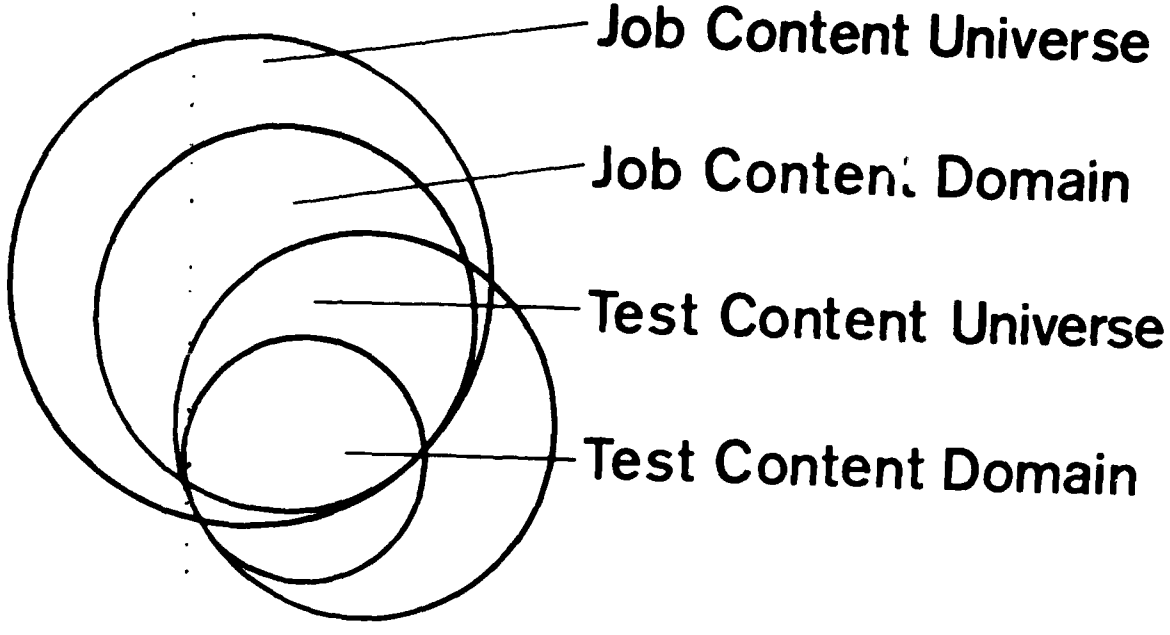


Figure 2. Venn diagrams relating universes and domains of job and test content

rarely if ever a component of the actual job content domain. Moreover, the scoring system is subject to contamination, just as is the scoring of any other test. That is, the obtained score an individual makes may reflect the attribute one wishes to infer, ability to do the job, but it may also reflect a variety of contaminations such as anxiety, ability to comprehend the verbal instructions, or perceptual skills in seeing cues for scoring enabling perceptive or test-wise people to make better scores than others.

All of this has a familiar ring after the earlier discussion of construct validity. All of the other possible components of a score represent the contaminations which construct validation, in its commitment to disconfirmatory research, is designed to investigate. To repeat: content validity is a special case of construct validity (Messick, 1975; Tenopyr, 1977).

ACCEPTANCE OF OPERATIONAL DEFINITIONS

> "Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few. Test validation, in fact, is widely regarded as the least satisfactory aspect of test development.... It is the purpose of this paper to develop an alternative explanation of the problem, and to propose an alternative solution. The basic difficulty in validating many tests arises, we believe, not from inadequate criteria but from logical and operational limitations of the concept of validity itself. We are persuaded that faster progress will be made toward better educational and psychological tests if validity is given a much more specific and restricted definition than is usually the case, and if it is no longer regarded as the supremely important quality of a remittal test" (Ebel, 1961, p. 640).

With these words, Ebel began a critique of the concept of validity as a major basis for evaluating tests. Many of the comments made in that paper are still highly applicable; people still tend to think of validity in terms of "real" traits, they still accept criterion measures

that have little if anything to do with the attributes being measured (and do not recognize that in doing so they have formed an external hypothesis), and the concept of validity is still far too broad to have scientific utility. Alternatives include the evaluation of reliability, normative data, the importance of the knowledge or abilities required by a test, convenience in the use of the test, and, most of all, meaningfulness.

Meaningfulness was also the primary yardstick for evaluation proposed by Messick (1975); his concept of meaningfulness, however, turns out to be equivalent to the concept of construct validity. Ebel, but not Messick, would evaluate a test simply as an operational definition of an attribute to be measured; the operations provide the meaning.

This writer takes the position that operational definitions of the attributes to be measured can, under certain circumstances, provide both a necessary and a sufficient evaluation of the scores obtained by using it; that is, under certain circumstances, no statement of validity is needed. It is operationalism, not validation, that provides the meaning for fundamental measurement of physical properties of length, time, or weight. As pointed out in the taxonomy of measurement, for measurement of such variables as these, one asks not whether the measurements are valid but whether they are accurate.

Some psychological measurement can also be defended as meaningful because of the operations involved in the measurement without recourse to the psychology's unique demand for validating the inferences from the scores. Operationalism does not always eliminate concern for validating inferences; in fact, it is sufficient only in relatively restricted cases (Ebel, 1956, 1961; Tenopyr, 1977). In Tenopyr's terms, there are some constructs for which the content of the measurement, i.e., the operational definition, is a sufficient evaluation. With reference to the taxonomy of variables describing attributes of people, it would appear that these

constructs would include certain physical attributes, psychomotor skills, task proficiencies, and, with a caveat, measures of job knowledge. (The caveat is that scores on job knowledge tests may be unduly influenced by reading abilities having little to do with the actual level of knowledge.)

For these constructs, at least in part, it would seem possible to evaluate the job relevance and meaningfulness of a personnel testing program on the basis of the operations alone. In a combination of two other publications (Guion, 1977, in press), the writer has presented a list of six requirements which, if met, constitute a sufficient evaluation of the use of a test so that issues of validity need not arise. With some modifications to fit the present context, and with emphasis on personnel testing and judgment of job relatedness, these will be reproduced here.

First, the content domain must consist of behavior the meaning of which is generally accepted. At the risk of sounding like Gertrude Stein, we can say that doing something (like driving a car) is generally accepted as evidence of the ability to do it. If a person reads a passage, it means that he can read the passage; if he does not read the passage, it may not mean an inability to read it (Messick, 1975), but it certainly means that he did not. In such examples, the meaning of the behavior is obvious; it requires no great inferential leap to interpret or to draw inferences from the behavior samples.

Second, both the test content domain and the job content domain should be unambiguously defined. The domains should be defined well enough that people who disagree on the definition can nevertheless agree on whether a particular task or statement or item belongs in or out of the domain. In the present age of litigation, agreements on the definition of a content domain are always tenuous. The amount of agreement needed does not depend on nailing down, in very precise language, every conceivable component of a domain. It is enough that the boundaries of the domain are sufficiently well established for agreement among reasonable and knowledgeable people.

Third, the test content domain must be relevant to the job content domain. The question of relevance is again a matter of judgment, and judgment requires some evidence of agreement. In originally presenting this third condition, the lack of a measure of the degree of agreement of the domains was deplored; it now seems that the extent of agreement among qualified judges that the two are comparable is sufficient.

Fourth, qualified judges must also agree that the test content domain has been adequately sampled. The need to define what is meant by qualified judges is particularly strong in this condition. From the point of view of personnel testing, the best qualified judges are usually people who have done the job in question or who have supervised the performance of that job. The required level of agreement would appear to be minimally that necessary to avoid conflict. Disagreements differ qualitatively. Some qualified judges will disagree on semantic grounds; others may disagree because of fundamental differences in value systems. The disagreement between plaintiff and defendant is a serious level of disagreement; the disagreement between one who would suggest a slight change in wording and one who prefers the existing wording is not a profound disagreement and need not be taken seriously in evaluating domain sampling. The question, therefore, is whether there is a consensus (a majority view) and whether there is a reasonable freedom from dissatisfaction with the consensus on the part of most qualified judges. This requirement holds for defining the boundaries of a content domain, for judging the relevance of a test content domain to a job content domain, and for judging the adequacy of the sampling of the test content domain.

Fifth, the response portion of the testing must be reliably observed and evaluated. In the original presentation of this point, it was said, "This does not refer to internal consistency, of course" (Guion, 1977, p. 7). The phrase "of course" is now regretted. At the very least, any measurement should have some degree of functional unity; if there is not even enough internal consistency for significant correlations to exist

between the component parts of a content sample, then the score of the content sample should be subdivided into reasonably internally consistent components. This comment, it should be pointed out, is a necessary consequence of saying that what has passed for content validity is in fact a special case of construct validity; the first requirement of construct validity is internal consistency.

A more important implication of this requirement is that observers who record observations must agree reasonably well on what they have seen. If the behavior to be observed is not defined well enough to permit inter-observer agreement, it violates the first condition of an operational definition based on content sampling.

Sixth, the method of scoring the content sample must be generally accepted by qualified judges as relatively free from contaminants reflecting irrelevant attributes of examinees or attributes of observers or materials. This implies no stringent demands for agreement among the judges. If there is a serious suggestion of contamination from judges who have made the previous judgments, some study inquiring into the construct validity of the scores may be necessary.

Intrinsic Validity. A different approach to operationalism can be drawn from a parallel to the concept of intrinsic validity (Gulliksen, 1950), another way in which the meaningfulness of an operational definition can be known by its outcomes. For example, if an examinee is coached to take the test, and coaching for the test improves both test performance and performance on the job, then scores on the test are intrinsically related to performance on the job. The investigation of this relationship is, of course, an empirical investigation; it does not rest upon the consensus of qualified judges. Nevertheless, it is only remotely related to its closest cousin among the validities, criterion-related validity. For the test to be accepted as an operational definition, under this heading, not only must a correlation between

test performance and job performance be obtained, but it must not be lost as a consequence of coaching.

Operationalism Based on Formal Structure. If work sample performance is to be evaluated by evaluating the product of that performance, and the product is a tangible object, then the measurement may consist of measuring weight, conductivity of solder connections, the amount of stress needed to break a weld, or similar physical measurement. Such measurements are formal, fundamental measurements and they need no justification by recourse to notions of validity.

The logic of formal measurement could be extended to some other areas of psychological measurement. Two possibilities seem worth mentioning which, if tests could be successfully constructed by these methods, would provide formal measurement that should be accepted without any concern for notions of validity. One of these uses Guttman scaling for content-referenced interpretations of scores; the other applies latent trait theory. These will be discussed in detail in the next section.

CHALLENGES TO CLASSICAL THEORY

Classical psychometric theory has its origin in the study of individual differences. This study requires maximum distinctions between individuals, that is, maximum variances within groups. All of classical theory is based upon variance and upon the subdivision of variance into systematic and error sources. A test is said to be reliable, for example, to the extent that the variance in a set of scores obtained through its use is free from random error of variance. In its broadest sense, validity is likewise defined as the extent to which the variance in a set of scores is relevant to the purposes of measurement. In test construction, the best items are those in which there is a good match between item variance and total test score variance. The unit of measurement in mental testing is the standard deviation, and the basis for interpreting

test scores is the relationship of one individual to another in distribution.

In short, the emphasis has been on relative measurement rather than on anything fundamental or absolute. The contributions of classical psychometric theory have been substantial, but they have led to some peculiar phenomena. For example, grades in a course of study such as physical education may be based not on the number of pushups one can do or the distance one can swim, but rather upon how many pushups or how many laps one can do in comparison to others in the class. Special characteristics of the class do not enter into the standard evaluation of performance using classical theory.

The illustration points out three objections that have been leveled against the use of classical psychometric theory for many forms of measurement in psychology in general and in personnel testing in particular:

1. The evaluation of measurement and of the interpretation of individual scores depends on the unique characteristics of the sample of people and the sample of items studied in the construction and standardization of the test.

2. Classical interpretations of scores provide no standard for the interpretation of an individual score beyond its relative position within the distribution of scores in the sample of people studied. If the distribution as a whole is quite high, a low score within that distribution is treated as a poor score, even if in some absolute term it would be considered high. Even the techniques for estimating true scores are based upon sample distribution; estimation of a so-called true score is simply a device for acknowledging the fallibility or unreliability of measurement. It does not take into account the relationship of that estimated true score to any standard of measurement.

3. Classical measurement theory offers no definition of the limits of the usefulness of the test or of the degree to which the classical statements of validity, reliability, or norms may be generalized. No sample is ever precisely like the sample upon which norm tables have been built, but those tables are

> consistently used for interpreting the scores of people not in that sample. To what extent do these interpretations apply to people who are different from the original sample in certain ways? To what extent can the standardized interpretations of scores as norms be applied to different sets of conditions? Such questions have no answer in classical psychometric theory.

Three challenges to classical psychometric theory can be identified and discussed as potential solutions to this set of problems: content-referenced measurement, latent trait theory, and generalizability theory. In addition, another "challenge" is based on the fact that psychometric theory evaluates only inferences from scores, not the effects of the uses of such inferences. Program evaluation will be briefly mentioned in this context.

CONTENT-REFERENCED MEASUREMENT

The term content-referenced measurement will be used here to apply to any measurement technique developed explicitly to interpret scores relative to some sort of standard. The nature of the standard may vary; it might be a relatively precise point, perhaps with very tight tolerances, as in measuring machined work products. It might be a much more diffuse range of measurements, as in defining a range of satisfactory "scores" in physiological measurements associated with health. It might be an arbitrary cutting point, above which some people are selected and others rejected. However it is defined (and the defining of a standard identifies one of the problems with the relevant literature), that definition results in interpretations of scores relative to the internal structure or content of the measuring instrument rather than to a distribution of obtained measures. Whatever it is, and there is much debate over its precise nature, the one point to be emphasized is that content-referenced measurement is not norm-referenced measurement!

In keeping with the APA Standards, the term chosen here is content-

referenced measurement in preference to the more common term, criterion-referenced measurement. In most problems in educational measurement, the distinction between the two terms may be trivial enough to explain why, despite the preference in the Standards, the former has not been adopted; moreover, the term, criterion-referenced, has been so widely accepted in educational circles that there is a very real problem in attempting to change it (Hambleton, Swaminathan, Algina, & Coulson, 1978). For personnel testing, however, the distinction is exceedingly important. The term criterion has been widely used to identify a variable external to the test itself. It is quite possible, particularly in the development of work sample tests, to construct the test so that scores on it can be directly interpreted in relation to a standard of job performance (criterion) measured externally. This may be more than simply using expectancy tables to interpret test scores, although that could be one example. It could also imply that a work sample constructed to abstract various components of the job can yield scores explicitly tied to such job performance measures as scrap rates or others. Such interpretation of scores in relation to external criteria has never been envisioned in the educational measurement literature on so-called criterion-referenced testing, but it is important enough in personnel testing to warrant special efforts to avoid confusion.

Moreover, the emphasis on content-referenced interpretation according to the Standards refers to those interpretations "where the score is directly interpreted in terms of performance at each point on the achievement continuum being measured" (APA et al., 1974, p. 19, emphasis added). This is clearly a different idea from much of the literature on criterion-referenced testing, which effectively treats any score in the distribution simply as above or below a specified score or standard.

In summary, content-referenced testing seems a preferable term because (a) it is more descriptive, (b) it avoids ambiguity, (c) it fits the terminology of the Standards, and (d) it avoids any implication of

dichotomy. The term is not the only one that might have been chosen. The relevant literature includes, in addition to content-referenced and criterion-referenced measurement, standards-referenced measurement, universe-referenced measurement, domain-referenced measurement, objective-referenced measurement, and mastery testing. Each of these terms has been proposed, and has its adherents, because of a special emphasis that is sought. This is a final advantage of the term chosen for this report, because it seems indebted to no prior bias.

The foregoing is more than a semantic exercise. The choice of language can influence substantially the directions taken in applying the diverse literature, some of which has been spawned less from an interest in making a new contribution to measurement theory than in challenging the old and established. The concept, under whatever name is chosen, has attracted very little attention among personnel testing specialists. Tenopyr (1977) said that "the notion of criterion-referenced test interpretation... has no application in an employment setting" (p. 51). Ebel (1977) seems to agree. The point of their rejection of the idea may be as much a rejection of the rhetoric leading to dichotomous scoring as of the idea of interpreting scores relative to a standard.

Certainly there are places in personnel testing where one should interpret measurement against some standard other than the mean of a distribution, even if it means a dichotomous interpretation. Certainly, where productivity is determined by the speed of a moving conveyor, the individual who cannot keep up with the conveyor belt is performing at an inadequate level, whether that person is at the bottom of a distribution or merely a standard deviation below the mean.

Work Samples as Content-Referenced Tests. Work samples constitute a special form of content-referenced testing; the principal evaluation of them is in terms of job relevance. The previous discussion of content domain sampling suggested that judgments of job-relatedness can be

simplified by thinking of a four-stage process of defining the most complete possible conception of the job (the job content universe), selecting a domain of interest from that universe, and then defining the related test content universe and domain.

A work sample test is developed by sampling from that final domain. In some cases, one might use work sample techniques to develop a test which is not strictly a sample of work performance but from which work performance might be inferred. It has become an accepted cliche for such tests to refer to "the inferential leap." Figure 3 is a whimsical attempt to show graphically (and perhaps whimsically) some limits to the appropriateness of the term.

Tests can be developed which literally sample job content adding only enough testing operations to provide a scoring system. Probationary assignments can be carefully chosen, and performance on them can be carefully evaluated. These are the most complete samples that can be developed for selection or certification purposes. Simulations represent, in varying degrees, abstractions of "real" job content; they are less precisely samples, shorter, and more standardized. Tests called "work samples" are usually also abstractions from job content, typically more abstract than simulations.

The meaning of abstraction, in this context, can be illustrated by referring again to the stenographer's job. In work sample testing, one does not try to create precisely every exact task and every exact environmental condition influencing task performance. Rather, one classifies various kinds of tasks (classification is itself a process of abstraction), and creates examples of the different classes; these, performed under standard conditions and scored according to rules which are not part of the job, become the work sample test. In all three cases, the performance evaluated is a direct sample of performance on the actual job. A small problem of inference may be introduced by the scoring or evaluation

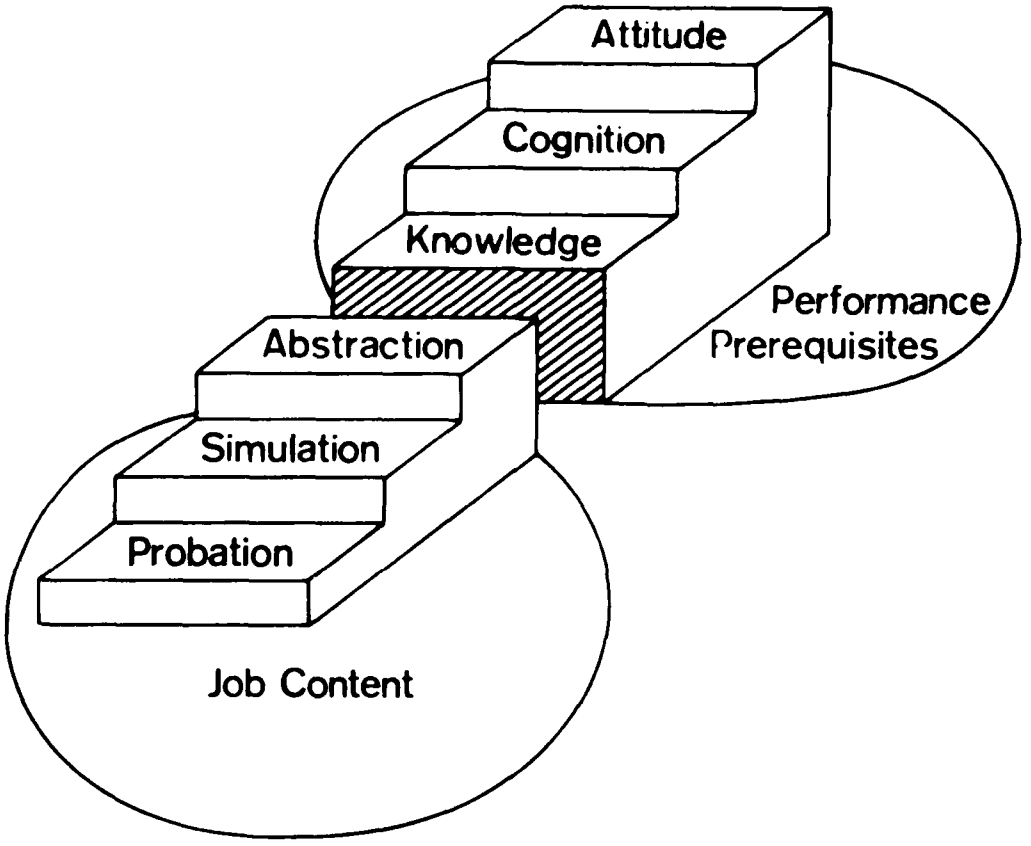


Figure 3. Samples and inferences in work sample testing

procedures, which can be contaminated by factors unrelated to real job performance, but the inference can hardly be said to require a leap.

A substantial portion of the job content domain, and therefore of an appropriate test content domain, consists of knowledge required to perform the job. In a work sample test consisting of tasks to be performed, the examinee gives evidence of the prerequisite knowledge by performing satisfactorily. In many certification programs, however, the work sample degenerates into a test of job knowledge alone. The verb has been chosen judiciously, for the job does not consist of knowledge isolated from action. (Some jobs consist primarily of knowledge. Where mastery of the knowledge component is likely to be a harder or more critical feature of the job than any actions using it, a job knowledge test is one kind of direct work sample.) The use of the job knowledge test usually implies the inference that having the knowledge leads to effective performance. Figure 3 suggests that this may not be a very great leap -- more an inferential step -- but that it is indeed more an inference than a sample. When one departs still further from actual performance of the job content, such as inferring prerequisite cognitive skills or essential attitudes, the measurement of these attributes really does require an inferential leap from test content to job content.

The greater the degree of abstraction from actual [illegible] job assignments, the more appropriate is the metaphor of the leap, and also the more appropriate is a criterion-related validation strategy. Work sample testing, if it is to be accepted on its own terms as content-referenced testing, should be concerned more with sampling than with inferring.

Job Analysis. Many kinds of job analysis procedures can be used for content-domain sampling. The procedures suggested here are illustrative, not prescriptive.

Briefly, the job analysis procedure may result in a series of formula

statements of the form, "__(Takes action)__ in __(setting)__ when __(action cue)__ occurs, using __(tools, knowledge, or skill)__." For a truck mechanic, such a statement might read, "Flushes truck radiator in garage when engine is said to overheat using water under pressure in flush tank." From such statements, one can specify what a worker does, what knowledge is necessary to do it, where information or material used in doing it comes from, and what happens after the task is finished. Such information defines the tasks, the methods, the prerequisites, and the contingencies that comprise the job content universe.

With the job content universe defined, panels of expert judges -- people who know the job well -- can whittle it down to a test content domain and can establish test specifications.

Assembling Test Content. In paper-and-pencil testing, one refers at this stage to writing items. The "items" in a work sample test might be tasks. Alternatively, tasks might be "subtests" and the "items" might be component characteristics of the process or product evaluated. In any event, scorable elements of the test are defined, developed, and assembled by experts on the job.

The essential meaning of the scores depend on the qualifications of the experts, the care with which they have reached the various judgments, and their overall degree of agreement. If all has been well done, scores (whether overall or on component tasks) can be interpreted directly with reference to the content of the test and without reference to any distribution of scores.

Scaling Test Content. Interpretation of scores with reference to test content can be facilitated and defended by establishing a formal metric for scoring. If a series of components of tasks, or components of a task content domain, can be arranged according to a genuine Guttman scale, all scores can be interpreted with reference to points on that

scale. This idea grows out of the illustration of "content standard scores" offered by Ebel (1962). In an arithmetic test of many items, ten items were selected. To this writer, merely glancing at the items identified, they seemed to fall along a scale of difficulty. If indeed they did fall along a scale, without overlapping discriminal dispersions, then any measurement technique using the other items could be tied statistically to the values along that scale. The result would be a content-referenced score with formal demonstration of transitivity.

An example of measurement approaching this sort of scaling is the Learning Assessment Program described by Grant and Bray (1970). In this program, examinees were given a series of tasks to learn to perform, seven in all; these were ordered so that it was necessary to have learned how to do task 1, to do task 2, and so on. The score for evaluating performance in this program was the level of the tasks learned. Thus, one who learned five tasks in a reasonable time was considered more proficient at the overall set of tasks than one who could only master three.

The same logic, it should be noted, can be applied to cognitive skill items. If it can be shown that a subset of items do form a reproducible scale, and if it can be further argued that these items constitute marker variables for a particular construct, then the formal properties of the scale should provide a sufficient operational definition for the evaluation of a testing program using it.

Evaluating Content-Referenced Tests. Do classical concepts of reliability and validity apply to content-referenced tests? Is it sensible to develop a test to measure, let us say, proficiency at the end of training (all trainees having at that time mastered the material of the test and therefore exhibiting no individual differences in proficiency), and to evaluate that test in classical terms defined on the basis of test score variance? Does it make sense to use norm-referenced concepts to evaluate content-referenced tests?

Much controversial literature has been devoted to such questions. The controversy probably stems from the non sequitur imbedded in the second question. It is indeed a non sequitur to equate measurement objectives with instructional objectives. A desire to have all trainees perform at an equally high level at the end of training is an objective demonstrably different from a desire to measure performance at that level. An analogy would be a Procrustean desire to stretch all little boys during their period of growth so that they can all be basketball players exactly seven feet tall. Success in the venture would lead to measures of height that have no variance; it does not follow that the yardstick used should be incapable of identifying other heights! Neither does it follow from recognizing this absurdity that variance-based statistics for determining reliability and validity are the appropriate evaluations.

In psychological measurement generally, validity has been an over-rated approach to evaluation; in work sample testing, validity concepts are far less important evaluations than are evaluations of job relevance. Content-referenced work samples developed according to the principles outlined above are assuredly job-related solely because of the method of their construction. Such a test, if scored with reference to a formal Guttman scale, could be evaluated particularly highly because of the meaningfulness of the metric. It is unfortunate that preoccupation with the concept of validity in classical measurement theory should make test users so willing to ignore the quality of measurement per se in their evaluations of the use of a test.

To assert that validity is an over-rated concept does not deny its real importance. In any sort of measurement where inferences are to be drawn beyond the descriptive character of the measuring instrument, the form of validity, generally called construct validity, is essential; nothing in content-referenced measurement relieves it of the obligation to be concerned over construct validity. Content domain sampling offers the first, and perhaps the only necessary, validity of inferences of

ability to do the job as sampled. If, however, the intended interpretation of the score seems to include something more than the test content (a frequent case), such as mastery or competence, then the score implies expectations the soundness of which must be demonstrated by the usual lines of evidence of construct validity (Linn, 1977).

That evidence may require experimental data showing that variance with groups judged as competent (or masters) is low relative to the variance between those groups and others judged as less competent (or nonmasters). Traditional validity coefficients may be useful, where obtained variances permit them, as results of inquiries into different aspects of the construct validity of the scores. Also, scores (or observations) on content-referenced tests must be reliably determined, although the nature of reliability may be conventional estimates of systematic variance, studies of the generalizability of scores, or the consistencies of classifications.

LATENT TRAIT THEORY

Under various names (latent structure analysis, item characteristic curve theory, Rasch model), latent trait theories constitute another approach to the construction of formal measuring instruments. The distinguishing importance of the method is that it defines item difficulties and other characteristics more or less independently of characteristics of the particular samples from which the data distributions are drawn.

Originally developed for the assessment of attitudes (Lazarsfeld, 1950), latent trait theory has subsequently been used mainly in the measurement of cognitive abilities (Lord & Novick, 1968; Hambleton & Cook, 1977). It can be used for at least some forms of work sample testing. Applications to tests of knowledge have been shown by Bejar, Weiss, & Gialluca (1977), and an application to personality measurement by Bejar (1977) seems directly applicable to measures of the quality of work sample products and other practical problems of personnel testing.

The Theoretical Foundations. Although the mathematical foundations of latent trait theory are beyond both the scope of this report and the abilities of the writer, a brief account of the nature of the theory is useful for discussions of its applicability.

An item characteristic curve can be identified in which the probability of a correct response to the item is seen as a function of the examinee's ability level. Various models exist for defining the function, one of which describes the item characteristic curve as a normal ogive. Figure 4 shows hypothetical item characteristic curves for three items. Item 1 has a fairly typical difficulty level; many people get it wrong, but many get it right. Item 2 is a very difficult item; only people of very high ability are likely to get it right, although people of low ability seem to get it right by guessing more often than on the other items. Item 3 is a highly discriminating item; most people with above average ability will get it right, and those with ability below average are unlikely to give a correct response.

Three parameters can be estimated for defining each of these curves. Parameter a is a discrimination index, proportional to the slope of the curve at the inflection point. Parameter b is a difficulty index, defined as the ability level on the base line corresponding to the point of inflection (the point corresponding to a .50 probability of correct response if the third parameter is zero). Parameter c is the probability of a correct response at infinitely low ability levels, often called the guessing parameter. Parameters estimated in a given analysis include the ability levels, identified as theta (θ) in Figure 4, of the people tested as well as the item parameters.

The theta scale may be defined arbitrarily in any given analysis; the numerical values of the difficulty parameters are therefore arbitrarily expressed for a given sample. However, parameters estimated from samples with different characteristics correlate very highly, even if one

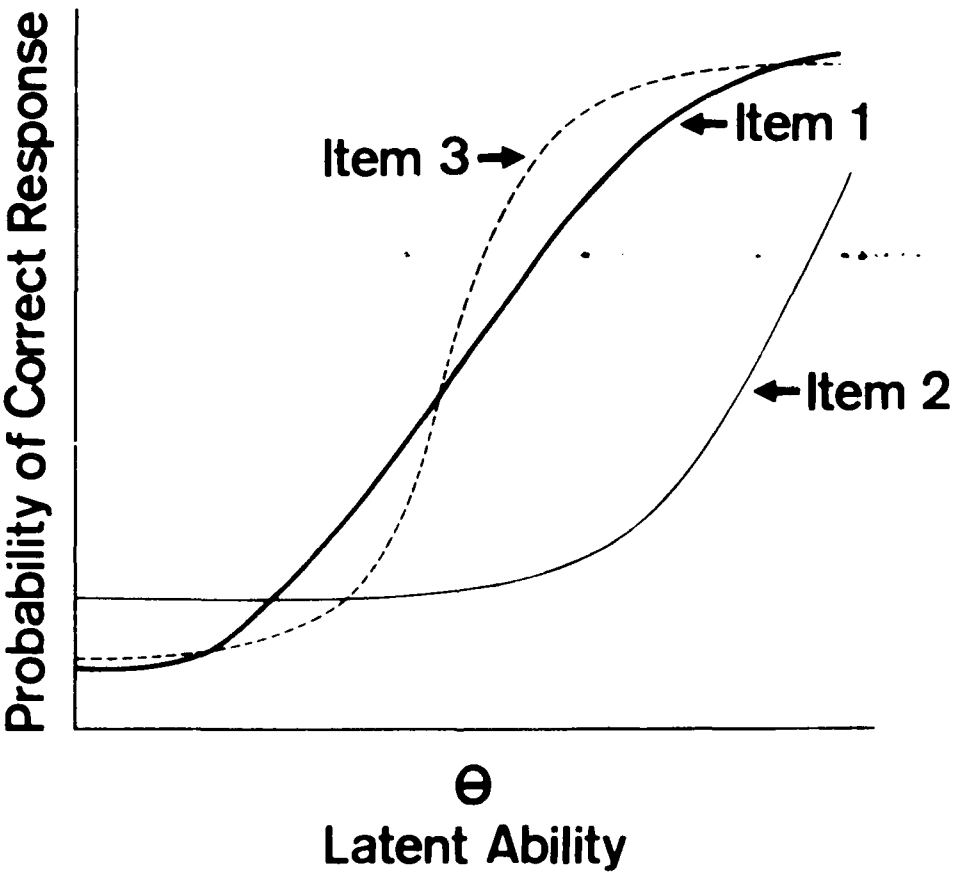


Figure 4. Item characteristic curves of three hypothetical items

sample consists of the low-scoring half of a distribution and the other sample consists of the high-scoring half (Rudner, 1976). Available equating procedures permit merging the latent ability scales for the population as a whole and expressing item characteristics in terms of that common scale. The resulting item characteristic curves are essentially congruent regardless of the sample from which they were developed. Failure to obtain such congruence indicates either a poor fit of the model or the possibility of item bias (Ironson, 1977).

The description presented here with Figure 1 (three parameters defining a normal ogive) refers to one of many models for latent trait analysis. There are logistic curves as well as normal ogives, and there are models that estimate only one or two of the parameters. The "two-parameter" models estimate discrimination values and difficulties; the "one-paramter" models estimate only difficulty levels. Multidimensional as well as unidimensional models have been proposed, and models are available for dichotomous, polychotomous, graded, or continuous responses (Samejima, 1969, 1972, 1973).

In classical psychometric theory, the standard error of measurement is generally treated as equal across the range of a distribution of scores. Its counterpart in latent trait analysis, the standard error of the estimate of ability, varies with the ability level. It is possible to construct item information curves showing the precision of the estimation of ability from responses on a single item at different ability levels. Tests measuring the same latent ability on the common scale can be assembled with different combinations of items, each with different item characteristic curves and item information curves. Combining item information curves across items yields a test information curve, the highest point of which is the level of ability at which the information (that is, the precision of estimating ability) provided by that set of items is greatest. Item characteristic curves may likewise be combined to yield a test characteristic curve in which the probability of an obtained

score is a function of the underlying ability level.

Uses of Latent Trait Analysis. If, for a given item, the item characteristic curves for two distinguishable groups of people are not essentially congruent, then that item cannot be said to be measuring the same latent ability in those two groups. Therefore, latent trait theory can be used to identify sources of item bias across race or sex groups.

This has implications for judgments about the adverse impact of tests used as decision tools. If there are substantial differences in obtained score distributions, the proportions of the groups selected (or classified into a desirable category) will differ. Current governmental regulations governing the use of employment procedures call for investigations to determine which of alternative selection tools will have lesser adverse effect, that is, which tests will have smaller mean differences in test performance.

If there are true subgroup differences, psychometric properties of the tests may affect the size of the adverse effect. Highly unreliable tests will have little adverse effect, for example. The problem can be highlighted by looking at test characteristic curves. The true differences in ability (as shown by the mean estimates of latent ability) are not influenced by the choice of test, but observed differences are. A test with a smaller slope on that curve will show less adverse effect than will a test with a characteristic curve that is steeper. In other words, even though the true differences are not changed by changing the test, the observed differences may be markedly greater for one test than for another -- and both can err in opposite directions. One of the tests may falsely exaggerate the true differences, while the other may falsely minimize them.

Working from item and test information curves, one can assemble small sets of items yielding the most precise possible ability estimates

at different ranges of ability levels (Lord, 1968; Weiss, 1974). If care has been taken to assure a full-range scale of ability in the development of an item bank, with known item characteristic curves, then any individual can be tested and located along that scale even using a unique set of items. Once the individual is located on that scale, the interpretation of his score is content referenced. For personnel testing, tests can be tailored not only for individuals but also for individual jobs requiring different levels of a particular ability, and standards for each job can be defined in terms of the ability levels appropriate.

Evaluation. Tests constructed using latent trait analysis can be evaluated with conventional concerns for job relatedness, reliability, and validity, but they may be better evaluated in other ways.

Job relatedness of work samples constructed by latent trait studies is no different from job relatedness of other work sample. In either case, it depends on the quality of the judgments made in defining the job content universe and in moving logically from that definition to a set of test specifications. Latent trait theory may, however, make it possible to develop abbreviated work samples that will be equally job related by identifying components that will maximize information at different levels of proficiency.

In latent trait theory, classical reliability is replaced by the idea of the information curve. Reliability coefficients can be manipulated by manipulating samples (Samejima, 1977); they are not sample-free. The standard error of measurement is a general statistic applying to all examinees in a distribution (or, if specially computed, in a specified broad range of the distribution). The standard error of estimate, however, is a value describing the precision of measurement at a particular point on the ability scale and is therefore far more informative. The test information curve gives evaluative information similar to that provided by reliability coefficients, but it does it better.

Construct validity is less important in latent trait studies than the fit of data to a model. If the data obtained from the items will indeed fit a latent trait model, they are certainly measuring something and doing so with internal consistency. Item construction proceeds, of course, in the context of a particular construct, so it is not difficult to define the underlying trait dimension. Construct validity, if of interest, is further assured if biased items (or items with other evidence of poor fit) are eliminated from the test or item pool as potential sources of contamination.

In general, however, validity statements are superfluous. The amount of research that goes into the development of such tests is indeed substantial. When that research has been completed, and measurement is expressed in terms of the underlying scale, that measurement is a sufficiently satisfactory operational definition of the construct being measured; no additional recourse to concepts of validity is necessary or informative.

GENERALIZABILITY THEORY

Generalizability theory (Cronbach, et al., 1972) does not challenge the norm-referenced basis of classical psychometric theory; it is, in fact, an extension of classical theory. The challenge it poses is the challenge to the undifferentiated distribution of error implicit in the classical formulation of true scores and error scores comprising an obtained score. Moreover, estimation of error in psychometric theory is built on the requirement of parallel tests, a condition not regularly satisfied in psychological measurement.

Any observed score is based on measurement obtained under a specified set of conditions. That set of conditions is but a sample of all of the possible sets that might have existed. Recognizing this, Cronbach and his associates ask investigators to define the universe of conditions, or the universe of possible observations, under which a person might be

tested. One generalizes in any actual use of tests from the sample to a universe of applicable conditions; generalizability studies make it possible to define the limits of possible generalizability for any test, a result particularly valuable in work sample testing.

An illustration of this implication may be helpful. Suppose that a work sample test is devised for measuring a specified skill at the end of training. Suppose, moreover, that the test is administered under traditional ideas of good test administration: good lighting, giving instructions carefully and consistently, special efforts to ascertain the reliability of observation, and a general effort to provide conditions optimally suited for maximizing performance of the examinee or reliability of the observations.

Now, no one is really interested specifically in how well the individual performs at the end of training except possibly the trainers. From an organizational point of view, the measurement of skill at the end of training is intended to generalize to conditions less optimal but more realistic, that is, to field rather than institutional settings. Obviously, there can be many different kinds of field conditions. Conditions can vary according to light sources, according to geographical climate, or according to variations in degrees of situational hostility.

A generalizability study, or a multiple facet analysis as it is also known, can be designed to determine the degree to which scores obtained in a sample measured under optimal conditions can be generalized to the different, non-optimal conditions of the study. Three possible kinds of findings can emerge: one may find that the inferences generalize quite well across conditions, one may find that they generalize not at all, or one may find that they will generalize to a limited subset of conditions, that is, that generalization across facets is possible only by the deletion of certain conditions.

One other point, too important for the possible implications of generalizability theory for work sample testing to be omitted from this brief discussion, is that the method permits one to estimate universe scores or expected obtained scores under specifiable combinations of facets. That is, even if there are substantial differences in performance under different sets of conditions, one may be able to generalize beyond the initial condition by making estimates of the obtained scores that would be expected under specified kinds of field conditions.

Program Evaluation. Alternatives to conventional validation procedures include evaluations of total programs using personnel tests. The use of assessment centers, in particular, has led to a situation in which the predictor is no longer a single test or small battery but the outcome of a complex assessment procedure expressed as the judgment of observers.

A less formal version of the same kind of thing occurs in an employment office where, instead of using a test and expectancy chart or cutting score, a series of assessment devices will be selected depending on the questions a decision-maker wishes to answer about a particular candidate for a particular job. Different batteries of tests may be used, different weights may be given to the same tests, and different questions may be asked. The procedure is frequently called clinical judgment or clinical prediction.

The total testing program, including judgments or decisions, can be evaluated in such circumstances if a quasi-experimental design can be used to compare the effectiveness of the performance, or work force stability, or other outcome in organization using the program to that in a different organization, reasonable well matched with the first, in which the program is not in use.

In a sense, this is criterion-related validation of the final judgment. It is, however, more in line with modern concerns for program

evaluation, and it is mentioned here as a potential stimulus to exploring the literature on program evaluation for its possible implications in the evaluation of personnel testing programs.

SUMMARY

Personnel testing programs have traditionally been evaluated in terms of the classical psychometric concepts of validity, particularly of criterion-related validity. The habit is well entrenched. Both the Standards (APA, et al., 1974) and the Principles for the Validation and Use of Personnel Selection Procedures (Division of Industrial-Organizational Psychology, 1975) give institutional support and encouragement to the habit. It is not a bad habit, like smoking, hazardous to the user's health and therefore to be broken; rather it is like eating, a habit to be tempered with moderation. Classical notions of validity have been valuable, but there are evaluative concepts that are more useful for some uses.

One of the difficulties with classical notions of validity is that there are too many of them and, in personnel testing, they have been forced to fit into too many Procrustean beds. The basic notion of validity as an evaluation of measurement has been stretched into something called content validity and squeezed into something else called criterion-related validity, neither of which refers to the quality or meaningfulness of measurement per se. Only investigations of construct validity provide useful insights into the meaning of measurement; what is called content validity is really better understood as content-oriented test development, and criterion-related validity is in reality the outcome of a test of a hypothesis.

In personnel testing, criterion-related validity holds a place of high honor. It is an established, useful approach for demonstrating the relationship of performance on the test to performance on the job --

a phrase which, when abbreviated, becomes job relatedness.

Job relatedness, or job relevance, is the most important single consideration in the evaluation of most personnel testing procedures, whether the testing is used to predict future performance, certify competency, evaluate performance, or validate some other variable. Criterion-related validity is a good source of evidence for judging the job relatedness of a test, but it is not the only one.

Equally important evidence of job relatedness is showing that the test is an acceptable operational definition of important aspects of job performance. Such a showing is based primarily on a thorough, rational process of getting information about a job and using expert opinion in defining domains, test specifications, and the relevance of individual items within the test. Surely, such information is at least on par with evidence of criterion-related validity serendipitously found using a test developed for a wide variety of general uses.

Another vitally important consideration in the evaluation of a test is the meaningfulness of scores obtained through its use. Meaningfulness can be established in part through the methods of establishing construct validity or from the methods of test construction. A very specific kind of meaning is derived through criterion-related studies. A quite different but perhaps equally valuable source of meaning is the concept of a latent trait.

In short, a score on a personnel test becomes meaningful in a variety of ways. It is meaningful if it can be interpreted in terms of a predicted level of future performance or of a probability of attaining some stated level of performance. It is meaningful if it can be interpreted as a proficiency measure on a sample of the actual job. It is meaningful if it can be interpreted directly in terms of a standard performance or in terms of a scale reflecting the variable being measured

without reference to an idiosyncratic distribution obtained from an available sample of people -- or, for that matter, of items. Its meaningfulness is enhanced to whatever degree it can be expressed as a score on a meaningful scale which retains that meaningfulness over a wide variety of circumstances. A content-referenced interpretation is at least as meaningful as a criterion-referenced interpretation (using the term here in its unusual sense of a score interpreted in terms of an external criterion variable). Thus methods of scaling or calibrating tests (such as latent trait analysis) need to be given a priority at least as high as that given to criterion-related validation in evaluating the meaningfulness of scores.

Classical test theory also evaluates tests in terms of reliability, meaning the freedom within a distribution of test scores from variance due to random error. Classical notions of reliability do not take systematic error into account. The application of the reliability concept to the evaluation of a single score is through the standard error of measurement, a value generally taken to be the same throughout the entire distribution of scores.

These are also useful evaluations, but they, too, can be improved upon through the use of newer ideas. Latent trait theory, for example, replaces the reliability theme with the idea of the information curve, using the standard error of estimated abilities as an index of precision at specific ability levels. Generalizability theory offers a much more comprehensive and useful accounting of various sources of error and their magnitudes, and it permits statements of both the limits of generalizability and the estimates of scores in different sets of conditions.

Modern measurement theory, although it has offered challenges to classical psychometric theory, has not reduced the usefulness of classical evaluations, especially in situations such as the use of tests or ratings in the measurement of such variables as attitude or personality

characteristics. For many other variables and for other methods of measurement, however, personnel testing needs to explore and exploit the possibilities of the newer theories. These possibilities are particularly relevant to work sample testing because it is most appropriately evaluated in terms of job relevance and its amenability to content-referenced interpretations of scores.

REFERENCES

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

Bejar, I. I. An application of the continuous response level model to personality measurement. Applied Psychological Measurement, 1977, 1, 509-521.

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of convocational and adaptive tests in the measurement of classroom achievement. (Resch. Rep. 77-7). Minneapolis: University of Minnesota, Psychometric Methods Program, 1977.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurement. New York: Wiley, 1972.

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

Division of Industrial-Organizational Psychology. Principles for the validation and use of personnel selection procedures. Dayton: Author, 1975.

Ebel, R. L. Obtaining and reporting evidence on content validity. Educational and Psychological Measurement, 1956, 16, 269-282.

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

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, R. L. Comments on some problems of employment testing. Personnel Psychology, 1977, 30, 55-63.

Grant, D. L., & Bray, D. W. Validation of employment tests for telephone company installation and repair occupations. Journal of Applied Psychology, 1970, 54, 7-14.

Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

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

Guion, R. M. Content validity -- the source of my discontent. Applied Psychological Measurement, 1977, 1, 1-10.

Guion, R. M. Scoring of content domain samples: The problem of fairness. Journal of Applied Psychology, in press.

Gulliksen, H. Intrinsic validity. American Psychologist, 1950, 5, 511-517.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

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

Ironson, G. H. A comparative study of several methods of assessing item bias. Unpublished doctoral dissertation, University of Wisconsin-Madison, 1977.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. Measurement and prediction. New York: Wiley, 1950.

Linn, R. L. Issues of validity in measurement for competency-based programs. Paper presented at the meeting of the National Council of Measurement in Education, New York, April, 1977.

Lord, F. M. Some test theory for tailor testing (ETS RB-68-38). Princeton, N.J.: Educational Testing Service, 1968.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

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

Rudner, L. M. Item and format bias and appropriateness. Washington, D.C.: Model Secondary School for the Deaf, 1976.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph No. 17, 1969.

Samejima, F. A general model for free-response data. Psychometrika Monograph No. 18, 1972.

Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38, 203-219.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247.

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

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.), Washington, D C.: American Council on Education, 1971.

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

Weiss, D. J. Strategies of adaptive ability measurement (Resch. Rep. 74-5). Minneapolis: University of Minnesota, Psychometric Methods Program, 1974.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on testing problems. Princeton, N.J.: Educational Testing Service, 1968.