A CRITICAL EXAMINATION OF THE CONCEPTS OF FACE VALIDITY

CHARLES I. MOSIER
Office of the Secretary of War[1]

FACE validity is a term that is bandied about in the field of test construction until it seems about to become a part of accepted terminology. The frequency of its use and the emotional reaction which it arouses—ranging almost from contempt to highest approbation—make it desirable to examine its meaning more closely. When a single term variously conveys high praise or strong condemnation, one suspects either ambiguity of meaning or contradictory postulates among those using the term. The tendency has been, I believe, to assume unaccepted premises rather than ambiguity, and beautiful friendships have been jeopardized when a chance remark about face validity has classed the speaker among the infidels.

An examination of the ways in which the term "face validity" has been used indicates three frequent meanings. These are sufficiently similar as to be confused, yet so different in their implications that to understand one meaning where another was intended leads to a wholly erroneous interpretation. This paper will analyze the various meanings which have been attributed to the term and it will then recommend that the term (and one of its meanings as well) be banished to outer darkness.

The three meanings which have been attributed to the term may be characterized as: (1) validity by *assumption,* (2) validity by *definition,* and (3) the *appearance* as well as the reality of validity. A fourth concept, validity by hypothesis, is closely related to the first two and deserves consideration in

[1] Opinions expressed in this paper are those of the author and do not necessarily reflect the policies of the War Department.

connection with them, although this concept has not generally been termed "face validity."

Validity by assumption: As used in this way, the term "face validity" carries the clear meaning that a test is assumed to be valid for the prediction of an external criterion if the items which compose it "appear on their face" to bear a common-sense relationship to the objective of the test. The assumption of validity in this case is asserted to be so strong that statistical evidence of validity is unnecessary; indeed, statistical evidence showing a lack of validity may be set aside by the strength of the assumption.

Validity by definition: For some tests, the objective is defined solely in terms of the population of questions from which the sample comprising the test was drawn, e.g., when the ability to handle the one hundred number facts of addition is tested by a sampling of those number facts. In these cases, the test is considered to be valid if the sample of items appears to the subject-matter expert to represent adequately the total universe of appropriate questions. The objective of the test is so defined that the index of reliability (the square root of the reliability coefficient) is, by definition, the measure of validity. This is so, not because of a definition of validity, but because of the way the objective of the test is defined. This situation is the one for which the term "face validity" was apparently coined.

Appearance of validity: In this usage, the term "face validity" implies that a test which is to be used in a practical situation should, in addition to having pragmatic or statistical validity, appear practical, pertinent and related to the purpose of the test as well; i.e., it should not only *be* valid but it should *also appear* valid. This usage of the term assumes that "face validity" is not validity in any usual sense of the word but merely an additional attribute of the test which is highly desirable in certain situations.

Validity by hypothesis: This concept, not generally associated with the term "face validity," is nevertheless sufficiently related to validity by asumption and validity by definition as to call for analysis at this point. The term "validity by hy-

pothesis" is used to characterize the following situation. Often, before the validity of a test can be empirically verified for a particular group by demonstration of its relationship to a satisfactory criterion, the test must be used to meet an immediate practical need. In such instances, the use of the test involves the hypothesis that it has a useful degree of validity. This hypothesis is based upon the designed similarity of the particular test to other tests already demonstrated to have known validity for the purpose in question. The validity of the test is not assumed in the sense that no further proof is required; neither is the objective of the test defined in such a way that the reliability of the test is evidence of its validity for the defined purpose. Rather the hypothesis is stated that, because of the sum total of previous knowledge relating to methods of predicting this particular criterion it is reasonable to suppose that a test of this sort will prove to be valid by the conventional statistical tests. This reasonable presumption, however, is subject to empirical verification by fact. Pending the opportunity for such verification, the presumption may be sufficiently strong as to justify the use of the test. Similarly, the physician studies the symptoms and the general condition of the patient and then, on the basis of his knowledge of the past effects of remedies upon similar symptoms in similar patients, prescribes treatment. He does this even though this combination of remedies has not occurred before in his experience and certainly not with this patient (who may have an unsuspected allergy which will defeat the purpose of the remedy).

With these four possible meanings of the term before us, it becomes profitable to examine each one in more detail.

Validity by Assumption

This conception of "face validity" is illustrated by the following quotations from a widely circulated testing handbook:

> Generally speaking, the validity of the test is best determined by using common sense in discovering that the test measures component abilities which exist both in the test situation and on the job. This common-sense approach to the problem of validity can be strengthened greatly by basing the estimate of the component of the job on a systematic observation of job analysis.

The term "face validity" is thus used to imply that the appearance of a relationship between the test and the external criterion is sufficient evidence of pragmatic validity. This use is a pernicious fallacy. This illegitimate usage has cast sufficient opprobrium on the term as to justify completely the recommendation that it be purged from the test technicians' vocabulary, even for its legitimate usage. The concept is the more dangerous because it is glib and comforting to those whose lack of time, resources, or competence prevent them from demonstrating validity (or invalidity) by any other method. Moreover, it is readily acceptable to the ordinary users of tests and its acceptance in these quarters lends the concept strength. This notion is also gratifying to the ego of the unwary test constructor. It implies that his knowledge and skill in the area of test construction are so great that he can unerringly design a test with the desired degree of effectiveness in predicting job success or in evaluating defined personality characteristics, and that he can do this so accurately that any further empirical verification is unnecessary. So strong is this ego complex that if statistical verification is sought and found lacking, the data represent something to be explained away by appeal to sampling errors or other convenient rationalization, rather than by scientific evidence which must be admitted into full consideration.

The concept of validity by assumption gains strength from the legitimate use of the term "face validity" to mean validity by definition. The superficial similarity, however, between the two concepts should not deceive us into accepting either the truth of the one or the necessary falsity of the other.

Any experienced test constructor can cite numerous instances of tests which appear so closely related to the external criterion that a high validity coefficient seems inevitable. The following example is to be considered merely one illustration which most readers can reproduce almost without limit from their own experience.

Two test construction agencies, each having a fairly large and competent staff, began work about the same time on an objective test to measure the clerical skills involved in alpha-

betical filing. Up to a certain point the two agencies worked independently, each devising its own test. Agency A, after an analysis of the job, constructed a test of which the following item is representative:

> "Below are five names, in random order. If the names were placed in strict alphabetical order, which name would be *third:* (1) John Meeder; (2) James Medway; (3) Thomas Madow; (4) Catherine Meagan; (5) Eleanor Meehan."

The second agency designed a test of skill in alphabetical filing in which the task was as follows:

> "In the following items you have one name which is underlined and four other names in alphabetical order. If you were to put the underlined name into the alphabetical series, indicate by the appropriate letter where it would go:
>
> Robert Carstens
>
> A._____
> Richard Carreton
> B._____
> Roland Casstar
> C._____
> Jack Corson
> D._____
> Edward Cranston
> E._____

There was a general agreement that each of these tests was face-valid and that each consisted of work-samples representative of the filing of alphabetical material. It was also agreed that if one were going to use two different tests to measure filing ability, it would be difficult to get two tests more closely similar than these and still have different tests. Had the concept of validity by assumption prevailed, there is little question that each test would have been considered highly valid.

An actual tryout, however, revealed quite different results from those expected. The correlation of the two tests in a sample of 43 clerical workers was .01, although the Kuder-Richardson reliabilities of the two tests were .81 and .89, respectively. We have here two tests which, on the basis of face validity by assumption, would be equally valid but which correlate substantially zero with one another. If one is valid, the other is not likely to be. What happens when the two tests

are studied, not for their correlation with each other, but for their correlation with what seems to be a reasonable criterion, namely supervisors' ratings of speed and accuracy in filing? For 72 employed workers where accuracy of filing materials was an important part of the job, the correlation between the first of the two tests described and the supervisors' rating was .09.[2] For the second test the correlation with the supervisors' ratings of accuracy in alphabetizing was .00. (That these results cannot be attributed to the unreliability of the supervisors' ratings is indicated by correlation coefficients of .40 and above between the same ratings and scores on other tests.) These two examples, therefore, as well as those which the reader's experience will readily bring to bear, are sufficient to demonstrate the fallacy involved in the statement that a test can be assumed to be valid without further verification if only it "measures component abilities which are judged by common sense to exist both in it and in the job."

Validity by Definition

The foregoing discussion has assumed an outside criterion measurable apart from the test itself. The discussion which follows is applicable rather to the situation, very frequent in educational measurement, in which the only available measure of the criterion (that which the test is intended to measure) is, because of the nature of the criterion, directly and intimately related to the test questions themselves. If the objective is the measurement of the pupils' skill in forming the elementary number combinations of addition, a test consisting of the one hundred possible combinations is presumably valid by definition. In this case the index of reliability can be taken as the validity coefficient. Even in this simple situation, the actual validity is limited by the reliability of the particular test, by the form in which the problems are presented, e.g., in words, in columns or in equations (e.g., four plus two equals ——; $+\,2$; $4+2=$ ——), the arrangement of the items and by the conditions of administration. As soon, however, as the test is

[2] The test did, however, show substantial correlation with other clerical skills and hence was useful in a general clerical battery, though not for its "face-valid" objective.

reduced from the totality of all situations which constitute the objective of measurement to a sample of those situations, the question recurs as to the extent to which the universe can be predicted from the sample. Moreover, it must be remembered that the relationship between test items and criterion behavior requires careful scrutiny. It is quite possible to design a test which apparently depends on the ability to perform the indicated additions, but is at the same time so dependent on verbal facility in understanding the directions, on speed of reaction, and on coding skills needed to record the answers, that the similarity between test situation and criterion situation is more apparent than real.

A further point which must be remembered in interpreting validity by definition is that it is frequently possible to establish several definitions of the criterion behavior, each obviously valid and yet each bearing far less than perfect relationship to the other. In the investigation of spelling ability, one obviously valid criterion of ability to spell might be the number of words correctly spelled from dictation. Should the words be dictated singly or in sentences, in a Brooklyn, Mobile, or Chicago accent? Another criterion which might be used, however, is a count of the number of words misspelled in compositions written by the pupils. Either of these criteria is, upon its face, a valid reflection of spelling ability. Nevertheless, empirical investigation is unlikely to show a perfect correlation between dictation and correct spelling in compositions, even after correction for attenuation. Which universe should be sampled to provide a face-valid test of spelling?

Finally, in the validation of a test by definition, it must be remembered that *the direction of the argument flows from the test to the definition of the criterion* rather than from the conceptually defined criterion to the test as a valid measure. The only proper statement which can be made about a test in terms of face validity by definition is that this test is a valid measure of that and only that universe of individual behavior patterns for which these items constitute a representative sample. If one is prepared to infer such a universe and consider *that* universe rather than one defined in any other way, such a concept

of validity may be useful. The necessity for inferring the conceptual nature of the universe from an examination of the sample still exists as a judgmental process and as one which is peculiarly subject to error.

If we return to the example of the two alphabetizing tests given in the section above we see how readily one may be misled into generalizing beyond the nature of the facts given. It is not difficult to draw the conclusion, from an inspection of the items, that these two tests were representative of the same universe and that therefore either test is a valid measure of the same set of skills. The fallacy of the conclusion, however, is attested by the absence of correlation between the two tests as cited above.

In educational achievement tests it is possible to outline the concepts to be covered in a particular course of study. These concepts may be sampled so systematically and so comprehensively that we are prepared to say the test questions constitute an adequate representation of all of the questions which might be asked on this course, in the light of its content and stated objectives. Even so, the questions may be so formulated that the crucial skills for achieving a high score on the examination are quite different from a knowledge of course content and the achievement of the stated objectives. We are correct in saying that the test is a valid measure of "whatever it measures reliably." We may be far from correct in inferring that the hypostatized "whatever" is what it appears to be on the face of the test. Nevertheless if we rely on validity by definition, we face the obligation of defining that "whatever" in some meaningful terms without running into the pitfall of *assuming* that the "whatever" is synonymous with the test constructor's objective in preparing the test.

As we examine critically the distinction between validity by assumption and validity by definition, we are led to see how tenuous is the dividing line between the scientifically defensible use, "validity by definition," and the totally unscientific and indefensible use, "validity by assumption."

Moreover, we do not escape the dilemma by refusing to recognize anything except external criteria. The validity of

the external criterion is just as much open to question as is the validity of the test which is being checked against it. Consider the situation in which a test purporting to measure clerical aptitude is "validated" by correlating test scores with salary (where salary is presumed to reflect the level of duties and responsibilities assigned). A high correlation between test score and salary level might well be taken, however, not as validation of the test but as validation of the agency's promotional system and an indication of the effectiveness with which the placement office had sought out and recommended for promotion the employees with the highest level of knowledge and skill. As Toops has pointed out, the criterion is a complex and elusive concept.[3] This paper is not the place for a systematic analysis of the nature of the criterion. It suffices to point out here that it is frequently possible to define in verbal, as distinct from operational, terms a criterion which is a socially significant independent measure of the behavior to be predicted by the test; such a definition is not in itself a sufficient guarantee that the criterion used to validate the test is itself valid.

The Appearance of Validity

In many situations it is highly desirable that the testing instrument should have a high degree of "consumer acceptance." These situations are most commonly found in, but by no means limited to, the field of employment testing. If a test is to be used effectively in achieving its objectives, it is essential that it actually be selected for use and that the results of the test be acceptable to those responsible for action on the basis of these results. In the area of public employment testing, e.g., civil service examining, the test must be acceptable not only to those using the test but to those taking the test as well. To a large extent this is also true in educational situations, particularly in the field of counseling. Up to a certain point the acceptability of the test can be carried by weight of authority. The board of examiners, the test technicians, or the counseling experts assert on the basis of their technical knowledge that the test is good, and their assertion is accepted without question.

[3] Toops, H. A. "The Criterion." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, IV (1944), 271–297.

In other situations, however, this assertion of authority is not sufficient to carry conviction. Moreover, the technical evidence on which such authoritative statements should be based is often neither comprehensible nor completely convincing to those who must be convinced.

In Civil Service situations, the candidate whose score is less than he expected is inclined to attribute his low score, not to his own deficiencies but to the impractical nature of the test in relation to the job for which he is being examined. His dissatisfaction with the test results and his feeling of injustice may, of course, have real merit. We have not yet reached the era of public personnel examining where all tests are technically sound. Whether or not there is merit in his claim, the legislature, the courts, and public opinion, the court of last appeal, are more readily impressed by superficial appearances than by correlation coefficients. It becomes highly important, therefore, that a test to be used in such a situation not only *be* valid in the pragmatic sense of affording reasonably accurate predictions of job competence, but *have the appearance of validity* as well.

This appearance of validity as an added attribute is important in terms of the acceptance of the test, not only by the persons being examined, but also by those operating officials who are charged with the responsibility for taking action based upon the test results. If sound tests are given and accurately reported, but the supervisor, interviewer, or counselor has no confidence in them, the results will not be used effectively.

In passing it should be noted that the concern of the Civil Service or merit system agency with the consumer acceptance of the test should not be merely a negative one of avoiding appeals or legislative pressures. In a democratic society the quality of public service is dependent to a large extent upon the public's opinion of the quality of public servants. If the examination by which public servants are selected (whether it be an objective test or an examination of the candidates' voting records) is such that competent persons in a particular occupation are convinced that they have no opportunity to demonstrate their competence, they will not file for the examination

or apply for the position. Since even the best Civil Service system can do no more than to select the best qualified persons of those who apply for positions, it is essential that every possible step be taken to insure that the most competent ones make application. They certainly will not do so if they believe that their examination will be impractical, theoretical, and deny them an opportunity to demonstrate their real ability. Moreover, in the face of such an attitude, statistical evidence on the validity of the test is likely to prove convincing only after an educational campaign extending to several generations of test-takers.

The foregoing discussion does not imply that predictive value is to be sacrificed to superficial appearances. Neither does it imply that a statistically valid test may be used only if it also has the appearance of practicality. It does imply, however, that the appearance of practicality is an objective sufficiently desirable in its own right that it may often be sought as an additional end consistent with the principal objective—predictive value.

The use of the term "face validity" to denote the appearance of a relationship to job performance as an attribute in addition to rather than instead of a statistical relationship, is frequently and unjustifiably confused with the notion of "face validity" by assumption. There is, however, a much clearer distinction between these two usages than between validity by assumption and validity by definition.

Validity by Hypothesis

This fourth view of validity has not, to the writer's knowledge, been explicitly termed face validity, although it contains certain elements of confusion with validity by assumption. In the construction of any test it is necessary to formulate certain hypotheses as to the most valid type and content to achieve a particular purpose. These hypotheses are held with a greater or less degree of confidence depending upon (a) the amount and the convincingness of available data showing that test items X have proved valid in situation Y, (b) the similarity of test item X to the proposed test items X′, and (c) the degree

of similarity between situation Y and situation Y′ in which the test is to be used. If the new test is very similar to one previously shown to be valid and if the new situation is very similar to that in which the test was valid, then we may proceed with a high degree of confidence that the proposed test will be valid in the situation in which it is to be used. This confidence, of course, never approaches certainty, and a verification of the hypothesis is always necessary.

Even though the questions and the methods of administration are identical for the two tests (if we may speak of two sets of identical questions as two tests), the measuring instrument will not be identical in its effect if its application has shifted from one group of subjects to another or from one testing situation to another. When a test has been adequately standardized on one population and found to be highly valid for the prediction of a particular skill in that population, the use of the same test for another population involves merely a hypothesis, rather than the certainty, of its validity as a measure of the same skill in the new situation. Even though we may have a high degree of confidence that the hypothesis will be confirmed, it is nevertheless a hypothesis. As we construct alternate forms of a test and apply them to new situations to predict the same set of skills, our degree of confidence becomes substantially less. The confidence level is also reduced when we use the same test to predict a somewhat different set of skills. For example, a test may be used to predict competence in clerical office work of a certain type in one agency when the test has been validated against proficiency in office work of a similar type but in another agency. In all these cases we are dealing with varying amounts of confidence in the validity of a test in a particular situation. The degree of confidence which justifies the use of an examining instrument in advance of its validation in the specific situation is a question of administrative judgment which is not wholly answerable by statistical techniques.

The foregoing discussion makes it clear that a validation study does not completely validate the test for use with another group of subjects but that it merely increases our confidence

that the test when applied to a group of "similar" subjects will prove similarly valid. Any selection of an existing test to serve a particular purpose (or construction of a new test to serve that purpose) therefore involves validity by hypothesis to a certain extent. The only situation in which we can escape the conclusion that our knowledge of the validity of a test is a hypothesis is the extremely limited one in which the test is validated on the identical subjects for which it is to be used administratively. Since validation of the test involves obtaining criterion measures (which are presumably superior to the test itself and would be used if it were not for the greater time and cost of securing them), the absurdity of using a test which has been prevalidated in this sense becomes immediately apparent. This does not lead, of course, to the absurd conclusion that a test may never be used; rather, it makes clear that when a test is used, its use is based upon a hypothesis in which we have more or less confidence depending upon the amount of research which has preceded its formulation. Our confidence in the test also depends upon the similarity between the research situation and the service-testing situation. Needless to say, this conclusion applies with equal force to all personnel evaluation and prediction devices.

It will be noted that validity by hypothesis departs from the concept of "face validity" in the preceding usages of the term. The first three usages discussed involve a superficial, common-sense similarity between test content and test objective. For example, in validity by assumption the similarity between test and job, without regard to statistical evidence of validity, is taken as sufficient. In validity by hypothesis, the similarity to a test for which there is statistical evidence of validity is tentatively accepted, without regard to its resemblance to the criterion. In validity by hypothesis, no such superficial similarity is assumed. On the basis of extensive previous research, one might legitimately propose that the ability to identify pictured hands as right hands or left hands would be a valid test for the prediction of the ability to read blueprints, although the superficial resemblance between the two tasks is slight. Nevertheless, certain controversies which

have been raised about face validity and the presumed necessity for prevalidating any test before it is used[4] make the discussion of validity by hypothesis appropriate in connection with the other uses of face validity.

Moreover, in validity by assumption, hypothesis, or definition, we are dealing with varying points on a continuum of degrees of certainty. In "assumption" we have, within the scientific frame of reference, no confidence whatever; in "hypothesis" we have varying degrees of confidence depending on the amount, quality and pertinence of the evidence from previous experience; in "definition," our confidence usually is greatest, but—and this must always be remembered—that confidence applies only to the trait or traits actually represented by the test items in relation to the sample and *not* to traits defined in any other way.

Summary and Conclusions

1. This paper has attempted an analysis of the various meanings of the term "face validity." These meanings, although superficially similar, lead to widely different conclusions.

2. The results of the analysis may be summarized as follows. Face validity is variously used to mean that:

a) The test bears a common-sense relationship to the measurement objective and therefore no statistical verification is necessary (*assumption*).

b) The test sets such a task that the universe of possible tasks (of which the test is a representative sample) is the only practicable criterion and the test is therefore a valid measure of the universe defined in terms of the sample. This implies merely that the test is a valid measure of whatever trait is measured reliably by the test (*definition*).

c) In the interests of the acceptability of the test to those most intimately concerned with its use, it is highly desirable that a test possess not only statistical validity,

[4] Strangely enough, many of those who insist upon the prevalidation of each written test continue to urge reliance upon other types of selection techniques which numerous research studies have almost unanimously shown to be without predictive value.

but also, as an added attribute, the appearance of practicality (*appearance*).

d) In the construction or selection of a particular test to be used for a particular objective with a particular group of subjects, recourse is always had to previous knowledge of the effectiveness of the same or similar tests applied to the same or similar subjects for the prediction of the same or similar attributes. On the basis of this previous research, the hypothesis is proposed that this test will be valid for the particular objective. The hypothesis is one which carries varying degrees of confidence: in some cases enough to justify the use of the test immediately, pending further investigation; in other cases so little confidence that such further investigation seems unprofitable. Even after there has been further investigation, however, we are left with a degree of confidence which is somewhat less than certainty, unless we are dealing with the same test, the same population and the same objectives (*hypothesis*).

3. Since the term "face validity" has become overlaid with a high degree of emotional content and since its referents are not only highly ambiguous but lead to widely divergent conclusions, it is recommended that the term be abandoned. Anyone intending to use the term should, instead, describe fully the *concept* which he originally intended to denote by "face validity." Even though writers may not always follow this recommendation, it is hoped that the foregoing analysis will prevent readers from drawing the improper conclusions that have frequently resulted from the indiscriminate uses made of the term in recent years.