CONSTRUCT VALIDITY IN PSYCHOLOGICAL TESTS

Lee J. Cronbach and Paul E. Meehl (1955)[1]

First published in *Psychological Bulletin, 52*, 281-302.

Validation of psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950-54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term *construct validity*.[2] This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers). The statements agreed upon by the Committee (and by committees of two other associations) were published in the *Technical Recommendations* (59). The present interpretation of construct validity is not "official" and deals with some areas where the Committee would probably not be unanimous. The present writers are solely responsible for this attempt to explain the concept and elaborate its implications.

Identification of construct validity was not an isolated development. Writers on validity during the preceding decade had shown a great deal of dissatisfaction with conventional notions of validity, and introduced new terms and ideas, but the resulting aggregation of types of validity seems only to have stirred the muddy waters. Portions of the distinctions we shall discuss are implicit in Jenkins' paper, "Validity for what?" (33), Gulliksen's "Intrinsic validity" (27), Goodenough's distinction between tests as "signs" and "samples" (22), Cronbach's separation of "logical" and "empirical" validity (11), Guilford's "factorial validity" (25), and Mosier's papers on "face validity" and "validity generalization" (49, 50). Helen Peak (52) comes close to an explicit statement of construct validity as we shall present it.

FOUR TYPES OF VALIDATION

The categories into which the *Recommendations* divide validity studies are: predictive validity, concurrent validity, content validity, and construct validity. The first two of these may be considered together as *criterion-oriented* validation procedures.

The pattern of a criterion-oriented [p. 282] study is familiar. The investigator is primarily interested in some criterion which he wishes to predict. He administers the test, obtains an independent criterion measure on the same subjects, and computes a correlation. If the criterion is obtained some time after the test is given, he is studying *predictive validity*. If

the test score and criterion score are determined at essentially the same time, he is studying *concurrent validity*. Concurrent validity is studied when one test is proposed as a substitute for another (for example, when a multiple-choice form of spelling test is substituted for taking dictation), or a test is shown to correlate with some contemporary criterion (e.g., psychiatric diagnosis).

Content validity is established by showing that the test items are a sample of a universe in which the investigator is interested. Content validity is ordinarily to be established deductively, by defining a universe of items and sampling systematically within this universe to establish the test.

Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not "operationally defined." The problem faced by the investigator is, "What constructs account for variance in test performance?" Construct validity calls for no new scientific approach. Much current research on tests of personality (9) is construct validation, usually without the benefit of a clear formulation of this process.

Construct validity is not to be identified solely by particular investigative procedures, but by the orientation of the investigator. Criterion-oriented validity, as Bechtoldt emphasizes (3, p. 1245), "involves the *acceptance* of a set of operations as an adequate definition of whatever is to be measured." When an investigator believes that no criterion available to him is fully valid, he perforce becomes interested in construct validity because this is the only way to avoid the "infinite frustration" of relating every criterion to some more ultimate standard (21). In content validation, *acceptance* of the universe of content as defining the variable to be measured is essential. Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured. Determining what psychological constructs account for test performance is desirable for almost any test. Thus, although the MMPI was originally established on the basis of empirical discrimination between patient groups and so-called normals (concurrent validity), continuing research has tried to provide a basis for describing the personality associated with each score pattern. Such interpretations permit the clinician to predict performance with respect to criteria which have not yet been employed in empirical validation studies (cf. 46, pp. 49-50, 110-111).

We can distinguish among the four types of validity by noting that each involves a different emphasis on the criterion. In predictive or concurrent validity, the criterion behavior is of concern to the tester, and he may have no concern whatsoever with the type of behavior exhibited in the test. (An employer does not care if a worker can manipulate blocks, but the score on the block test may predict something he cares about.) Content validity is studied when the tester *is* concerned with the type of behavior involved in the test performance. Indeed, if the test is a work sample, the behavior represented in the test may be an end in itself. Construct validity is ordinarily studied when the tester has no definite criterion measure of the quality with which he is concerned, and must use indirect measures. Here the trait or quality underlying the test if of central importance, rather than either the test behavior or the scores on the criteria (59, p. 14).

[p. 283] Construct validation is important at times for every sort of psychological test: aptitude, achievement, interests, and so on. Thurstone's statement is interesting in this connection:

In the field of intelligence tests, it used to be common to define validity as the correlation between a test score and some outside criterion. We have reached a stage of sophistication where the test-criterion correlation is too coarse. It is obsolete. If we attempted to ascertain the validity of a test for the second space-factor, for example, we would have to get judges [to] make reliable judgments about people as to this factor. Ordinarily their [the available judges'] ratings would be of no value as a criterion. Consequently, validity studies in the cognitive functions now depend on criteria of internal consistency . . . (60, p. 3).

Construct validity would be involved in answering such questions as: To what extent is this test culture-free? Does this test of "interpretation of data" measure reading ability, quantitative reasoning, or response sets? How does a person with A in Strong Accountant, and B in Strong CPA, differ from a person who has these scores reversed?

Example of construct validation procedure. Suppose measure *X* correlates .50 with *Y*, the amount of palmar sweating induced when we tell a student that he has failed a Psychology I exam. Predictive validity of *X* for *Y* is adequately described by the coefficient, and a statement of the experimental and sampling conditions. If someone were to ask, "Isn't there perhaps another way to interpret this correlation?" or "What other kinds of evidence can you bring to support your interpretation?", we would hardly understand what he was asking because no interpretation has been made. These questions become relevant when the correlation is advanced as evidence that "test *X* measures anxiety proneness." Alternative interpretations are possible; e.g., perhaps the test measures "academic aspiration," in which case we will expect different results if we induce palmar sweating by economic threat. It is then reasonable to inquire about other *kinds* of evidence.

Add these facts from further studies: Test *X* correlates .45 with fraternity brothers' ratings on "tenseness." Test *X* correlates .55 with amount of intellectual inefficiency induced by painful electric shock, and .68 with the Taylor Anxiety scale. Mean *X* score decreases among four diagnosed groups in this order: anxiety state, reactive depression, "normal," and psychopathic personality. And palmar sweat under threat of failure in Psychology I correlates .60 with threat of failure in mathematics. Negative results eliminate competing explanations of the *X* score; thus, findings of negligible correlations between *X* and social class, vocational aim, and value-orientation make it fairly safe to reject the suggestion that *X* measures "academic aspiration." We can have substantial confidence that *X* does measure anxiety proneness if the current theory of anxiety can embrace the variates which yield positive correlations, and does not predict correlations where we found none.

KINDS OF CONSTRUCTS

At this point we should indicate summarily what we mean by a construct, recognizing that much of the remainder of the paper deals with this question. A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct. We expect a

person at any time to possess or not possess a qualitative attribute (amnesia) or structure, or to possess some degree of a quantitative attrib- [p. 284] bute (cheerfulness). A construct has certain associated meanings carried in statements of this general character: Persons who possess this attribute will, in situation *X*, act in manner *Y* (with a stated probability). The logic of construct validation is invoked whether the construct is highly systematized or loose, used in ramified theory or a few simple propositions, used in absolute prepositions or probability statements. We seek to specify how one is to defend a proposed interpretation of a test: *we are not recommending any one type of interpretation*.

The constructs in which tests are to be interpreted are certainly not likely to be physiological. Most often they will be traits such as "latent hostility" or "variable in mood," or descriptions in terms of an educational objective, as "ability to plan experiments." For the benefit of readers who may have been influenced by certain eisegeses of MacCorquodale and Meehl (40), let us here emphasize: Whether or not an interpretation of a test's properties or relations involves questions of construct validity is to be decided by examining the entire body of evidence offered, together with what is asserted about the test in the context of this evidence. Proposed identifications of constructs allegedly measured by the test with constructs of other sciences (e.g., genetics, neuroanatomy, biochemistry) make up only *one* class of construct-validity claims, and a rather minor one at present. Space does not permit full analysis of the relation of the present paper to the MacCorquodale-Meehl distinction between hypothetical constructs and intervening variables. The philosophy of science pertinent to the present paper is set forth later in the section entitled, "The nomological network."

THE RELATION OF CONSTRUCTS TO "CRITERIA"

Critical View of the Criterion Implied

An unquestionable criterion may be found in a practical operation, or may be established as a consequence of an operational definition. Typically, however, the psychologist is unwilling to use the directly operational approach because he is interested in building theory about a generalized construct. A theorist trying to relate behavior to "hunger" almost certainly invests that term with meanings other than the operation "elapsed-time-since-feeding." If he is concerned with hunger as a tissue need, he will not accept time lapse as *equivalent* to his construct because it fails to consider, among other things, energy expenditure of the animal.

In some situations the criterion is no more valid than the test. Suppose, for example, that we want to know if counting the dots on Bender-Gestalt figure five indicates "compulsive rigidity," and take psychiatric ratings on this trait as a criterion. Even a conventional report on the resulting correlation will say something about the extent and intensity of the psychiatrist's contacts and should describe his qualifications (e.g., diplomate[*sic*] status? analyzed?).

Why report these facts? Because data are needed to indicate whether the criterion is any good. "Compulsive rigidity" is not really intended to mean "social stimulus value to psychiatrists." The implied trait involves a range of behavior-dispositions which may be

very imperfectly sampled by the psychiatrist. Suppose dot-counting does not occur in a particular patient and yet we find that the psychiatrist has rated him as "rigid." When questioned the psychiatrist tells us that the patient was a rather easy, free-wheeling sort; [p. 285] however, the patient *did* lean over to straighten out a skewed desk blotter, and this, viewed against certain other facts, tipped the scale in favor of a "rigid" rating. On the face of it, counting Bender dots may be just as good (or poor) a sample of the compulsive-rigidity domain as straightening desk blotters is.

Suppose, to extend our example, we have four tests on the "predictor" side, over against the psychiatrist's "criterion," and find generally positive correlations among the five variables. Surely it is artificial and arbitrary to impose the "test-should-perdict-criterion" pattern on such data. The psychiatrist samples verbal content, expressive pattern, voice, posture, etc. The psychologist samples verbal content, perception, expressive pattern, etc. Our proper conclusion is that, from this evidence, the four tests and the psychiatrist all assess some common factor.

The asymmetry between the "test" and the so-designated "criterion" arises only because the terminology of predictive validity has become a commonplace in test analysis. In this study where a construct is the central concern, any distinction between the merit of the test and criterion variables would be justified only if it had already been shown that the psychiatrist's theory and operations were excellent measures of the attribute.

INADEQUACY OF VALIDATION IN TERMS OF SPECIFIC CRITERIA

The proposal to validate constructual interpretations of tests runs counter to suggestions of some others. Spiker and McCandless (57) favor an operational approach. Validation is replaced by compiling statements as to how strongly the test predicts other observed variables of interest. To avoid requiring that each new variable be investigated completely by itself, they allow two variables to collapse into one whenever the properties of the operationally defined measures are the same: "If a new test is demonstrated to predict the scores on an older, well-established test, then an evaluation of the predictive power of the older test may be used for the new one." But accurate inferences are possible only if the two tests correlate so highly that there is negligible reliable variance in either test, independent of the other. Where the correspondence is less close, one must either retain all the separate variables operationally defined or embark on construct validation.

The practical user of tests must rely on constructs of some generality to make predictions about new situations. Test *X* could be used to predict palmar sweating in the face of failure without invoking any construct, but a counselor is more likely to be asked to forecast behavior in diverse or even unique situations for which the correlation of test *X* is unknown. Significant predictions rely on knowledge accumulated around the generalized construct of anxiety. The *Technical Recommendations* state:

It is ordinarily necessary to evaluate construct validity by integrating evidence from many different sources. The problem of construct validation becomes especially acute in the clinical field since for many of the constructs dealt with it is not a question of finding an imperfect criterion but of finding any criterion at all. The psychologist interested in

construct validity for clinical devices is concerned with making an estimate of a hypothetical internal process, factor, system, structure, or state and cannot expect to find a clear unitary behavioral criterion. An attempt to identify any one criterion measure or any composite as *the* criterion aimed at is, however, usually unwarranted (59, p. 14-15).

This appears to conflict with arguments for specific criteria prominent at places in the testing literature. [p. 286] Thus Anastasi (2) makes many statements of the latter character: "It is only as a measure of a specifically defined criterion that a test can be objectively validated at all . . . To claim that a test measures anything over and above its criterion is pure speculation" (p. 67). Yet elsewhere this article supports construct validation. Tests can be profitably interpreted if we "know the relationships between the tested behavior . . . and other behavior samples, none of these behavior samples necessarily occupying the preeminent position of a criterion" (p. 75). Factor analysis with several partial criteria might be used to study whether a test measures a postulated "general learning ability." If the data demonstrate specificity of ability instead, such specificity is "useful in its own right in advancing our knowledge of behavior; it should not be construed as a weakness of the tests" (p. 75).

We depart from Anastasi at two points. She writes, "The validity of a psychological test should not be confused with an analysis of the factors which determine the behavior under consideration." We, however, regard such analysis as a most important type of validation. Second, she refers to "the will-o'-the-wisp of psychological processes which are distinct from performance" (2, p. 77). While we agree that psychological processes are elusive, we are sympathetic to attempts to formulate and clarify constructs which are evidenced by performance but distinct from it. Surely an inductive inference based on a pattern of correlations cannot be dismissed as "pure speculation."

Specific Criteria Used Temporarily: The "Bootstraps" Effect

Even when a test is constructed on the basis of a specific criterion, it may ultimately be judged to have greater construct validity than the criterion. We start with a vague concept which we associate with certain observations. We then discover empirically that these observations covary with some other observation which possesses greater reliability or is more intimately correlated with relevant experimental changes than is the original measure, or both. For example, the notion of temperature arises because some objects feel hotter to the touch than others. The expansion of a mercury column does not have face validity as an index of hotness. But it turns out that (*a*) there is a statistical relation between expansion and sensed temperature; (*b*) observers employ the mercury method with good interobserver agreement; (*c*) the regularity of observed relations is increased by using the thermometer (e.g., melting points of samples of the same material vary little on the thermometer; we obtain nearly linear relations between mercury measures and pressure of a gas). Finally, (*d*) a theoretical structure involving unobservable microevents -- the kinetic theory -- is worked out which explains the relation of mercury expansion to heat. This whole process of conceptual enrichment begins with what in retrospect we see as an extremely fallible "criterion" -- the human temperature sense. That original criterion has now been relegated to a peripheral position. We have lifted ourselves by our bootstraps, but in a legitimate and fruitful way.

Similarly, the Binet scale was first valued because children's scores tended to agree with judgments by schoolteachers. If it had not shown this agreement, it would have been discarded along with reaction time and the other measures of ability previously tried. Teacher judgments once constituted the criterion against [p. 287] which the individual intelligence test was validated. But if today a child's IQ is 135 and three of his teachers complain about how stupid he is, we do not conclude that the test has failed. Quite to the contrary, if no error in test procedure can be argued, we treat the test score as a valid statement about an important quality, and define our task as that of finding out what other variables -- personality, study skills, etc. -- modify achievement or distort teacher judgment.

EXPERIMENTATION TO INVESTIGATE CONSTRUCT VALIDITY

Validation Procedures

We can use many methods in construct validation. Attention should particularly be drawn to Macfarlane's survey of these methods as they apply to projective devices (41).

Group differences. If our understanding of a construct leads us to expect two groups to differ on the test, this expectation may be tested directly. Thus Thurstone and Chave validated the Scale for Measuring Attitude Toward the Church by showing score differences between church members and nonchurchgoers. Churchgoing is not *the* criterion of attitude, for the purpose of the test is to measure something other than the crude sociological fact of church attendance; on the other hand, failure to find a difference would have seriously challenged the test.

Only coarse correspondence between test and group designation is expected. Too great a correspondence between the two would indicate that the test is to some degree invalid, because members of the groups are expected to overlap on the test. Intelligence test items are selected initially on the basis of a correspondence to age, but an item that correlates .95 with age in an elementary school sample would surely be suspect.

Correlation matrices and factor analysis. If two tests are presumed to measure the same construct, a correlation between them is predicted. (An exception is noted where some second attribute has positive loading in the first test and negative loading in the second test; then a low correlation is expected. This is a testable interpretation provided an external measure of either the first or the second variable exists.) If the obtained correlation departs from the expectation, however, there is no way to know whether the fault lies in test A, test B, or the formulation of the construct. A matrix of intercorrelations often points out profitable ways of dividing the construct into more meaningful parts, factor analysis being a useful computational method in such studies.

Guilford (26) has discussed the place of factor analysis in construct validation. His statements may be extracted as follows:

"The personnel psychologist wishes to know 'why his tests are valid.' He can place tests and practical criteria in a matrix and factor it to identify 'real dimensions of human personality.' A factorial description is exact and stable; it is economical in explanation; it leads to the

creation of pure tests which can be combined to predict complex behaviors." It is clear that factors here function as constructs. Eysenck, in his "criterion analysis" (18), goes farther than Guilford, and shows that factoring can be used explicitly to test hypotheses about constructs.

Factors may or may not be weighted with surplus meaning. Certainly when they are regarded as "real dimensions" a great deal of surplus meaning is implied, and the interpreter must shoulder a substan- [p. 288] tial burden of proof. The alternative view is to regard factors as defining a working reference frame, located in a convenient manner in the "space" defined by all behaviors of a given type. Which set of factors from a given matrix is "most useful" will depend partly on predilections, but in essence the best construct is the one around which we can build the greatest number of inferences, in the most direct fashion.

Studies of internal structure. For many constructs, evidence of homogeneity within the test is relevant in judging validity. If a trait such as *dominance* is hypothesized, and the items inquire about behaviors subsumed under this label, then the hypothesis appears to require that these items be generally intercorrelated. Even low correlations, if consistent, would support the argument that people may be fruitfully described in terms of a generalized tendency to dominate or not dominate. The general quality would have power to predict behavior in a variety of situations represented by the specific items. Item-test correlations and certain reliability formulas describe internal consistency.

It is unwise to list uninterpreted data of this sort under the heading "validity" in test manuals, as some authors have done. High internal consistency may *lower* validity. Only if the underlying theory of the trait being measured calls for high item intercorrelations do the correlations support construct validity. Negative item-test correlations may support construct validity, provided that the items with negative correlations are believed irrelevant to the postulated construct and serve as suppressor variables (31, p. 431-436; 44).

Study of distinctive subgroups of items within a test may set an upper limit to construct validity by showing that irrelevant elements influence scores. Thus a study of the PMA space tests shows that variance can be partially accounted for by a response set, tendency to mark many figures as similar (12). An internal factor analysis of the PEA Interpretation of Data Test shows that in addition to measuring reasoning skills, the test score is strongly influenced by a tendency to say "probably true" rather than "certainly true," regardless of item content (17). On the other hand, a study of item groupings in the DAT Mechanical Comprehension Test permitted rejection of the hypothesis that knowledge about specific topics such as gears made a substantial contribution to scores (13).

Studies of change over occasions. The stability of test scores ("retest reliability," Cattell's "N-technique") may be relevant to construct validation. Whether a high degree of stability is encouraging or discouraging for the proposed interpretation depends upon the theory defining the construct.

More powerful than the retest after uncontrolled intervening experiences is the retest with experimental intervention. If a transient influence swings test scores over a wide range,

there are definite limits on the extent to which a test result can be interpreted as reflecting the typical behavior of the individual. These are examples of experiments which have indicated upper limits to test validity: studies of differences associated with the examiner in projective testing, of change of score under alternative directions ("tell the truth" vs. "make yourself look good to an employer"), and of coachability of mental tests. We may recall Gulliksen's distinction (27): When the coaching is of a sort that improves the pupil's intellectual functioning in [p. 289] school, the test which is affected by the coaching has validity as a measure of intellectual functioning; if the coaching improves test taking but not school performance, the test which responds to the coaching has poor validity as a measure of this construct.

Sometimes, where differences between individuals are difficult to assess by any means other than the test, the experimenter validates by determining whether the test can detect induced intra-individual differences. One might hypothesize that the Zeigarnik effect is a measure of ego involvement, i.e., that with ego involvement there is more recall of incomplete tasks. To support such an interpretation, the investigator will try to induce ego involvement on some task by appropriate directions and compare subjects' recall with their recall for tasks where there was a contrary induction. Sometimes the intervention is drastic. Porteus finds (53) that brain-operated patients show disruption of performance on his maze, but do not show impaired performance on conventional verbal tests and argues therefrom that his test is a better measure of planfulness.

Studies of process. One of the best ways of determining informally what accounts for variability on a test is the observation of the person's process of performance. If it is supposed, for example, that a test measures mathematical competence, and yet observation of students' errors shows that erroneous reading of the question is common, the implications of a low score are altered. Lucas in this way showed that the Navy Relative Movement Test, an aptitude test, actually involved two different abilities: spatial visualization and mathematical reasoning (39).

Mathematical analysis of scoring procedures may provide important negative evidence on construct validity. A recent analysis of "empathy" tests is perhaps worth citing (14). "Empathy" has been operationally defined in many studies by the ability of a judge to predict what responses will be given on some questionnaire by a subject he has observed briefly. A mathematical argument has shown, however, that the scores depend on several attributes of the judge which enter into his perception of *any* individual, and that they therefore cannot be interpreted as evidence of his ability to interpret cues offered by particular others, or his intuition.

The Numerical Estimate of Construct Validity

There is an understandable tendency to seek a "construct validity coefficient." A numerical statement of the degree of construct validity would be a statement of the proportion of the test score variance that is attributable to the construct variable. This numerical estimate can sometimes be arrived at by a factor analysis, but since present methods of factor analysis are based on linear relations, more general methods will ultimately be needed to deal with many quantitative problems of construct validation.

Rarely will it be possible to estimate definite "construct saturations," because no factor corresponding closely to the construct will be available. One can only hope to set upper and lower bounds to the "loading." If "creativity" is defined as something independent of knowledge, then a correlation of .40 between a presumed test of creativity and a test of arithmetic knowledge would indicate that at least 16 per cent of the reliable test variance is irrelevant to creativity as defined. Laboratory performance on problems such as Maier's "hatrack" would scarcely be [p. 290] an ideal measure of creativity, but it would be somewhat relevant. If its correlation with the test is .60, this permits a tentative estimate of 36 per cent as a lower bound. (The estimate is tentative because the test might overlap with the irrelevant portion of the laboratory measure.) The saturation seems to lie between 36 and 84 per cent; a cumulation of studies would provide better limits.

It should be particularly noted that rejecting the null hypothesis does not finish the job of construct validation (35, p. 284). The problem is not to conclude that the test "is valid" for measuring the construct variable. The task is to state as definitely as possible the degree of validity the test is presumed to have.

THE LOGIC OF CONSTRUCT VALIDATION

Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim. The philosophy of science which we believe does most justice to actual scientific practice will not be briefly and dogmatically set forth. Readers interested in further study of the philosophical underpinning are referred to the works by Braithwaite (6, especially Chapter III), Carnap (7; 8, pp. 56-69), Pap (51), Sellars (55, 56), Feigl (19, 20), Beck (4), Kneale (37, pp. 92-110), Hempel (29; 30, Sec. 7).

The Nomological Net

The fundamental principles are these:

1. Scientifically speaking, to "make clear what something *is*" means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a *nomological network*.

2. The laws in a nomological network may relate (*a*) observable properties or quantities to each other; or (*b*) theoretical constructs to observables; or (*c*) different theoretical constructs to one another. These "laws" may be statistical or deterministic.

3. A necessary condition for a construct to be scientifically admissible is that it occur in a nomological net, at least *some* of whose laws involve observables. Admissible constructs may be remote from observation, i.e., a long derivation may intervene between the nomologicals which implicitly define the construct, and the (derived) nomologicals of type *a*. These latter propositions permit predictions about events. The construct is not "reduced" to the observations, but only combined with other constructs in the net to make predictions about observables.

4. "Learning more about" a theoretical construct is a matter of elaborating the nomological network in which it occurs, or of increasing the definiteness of the components. At least in the early history of a construct the network will be limited, and the construct will as yet have few connections.

5. An enrichment of the net such as adding a construct or a relation to theory is justified if it generates nomologicals that are confirmed by observation or if it reduces the number of nomologicals required to predict the same observations. When observations will not fit into the network as it stands, the scientist has a certain freedom in selecting where to modify the network. That is, there may be alternative constructs or ways of organizing the net which for the time being are equally defensible.

6. We can say that "operations" [p. 291] which are qualitatively very different "overlap" or "measure the same thing" if their positions in the nomological net tie them to the same construct variable. Our confidence in this identification depends upon the amount of inductive support we have for the regions of the net involved. It is not necessary that a direct observational comparison of the two operations be made -- we may be content with an intranetwork proof indicating that the two operations yield estimates of the same network-defined quantity. Thus, physicists are content to speak of the "temperature" of the sun and the "temperature" of a gas at room temperature even though the test operations are nonoverlapping because this identification makes theoretical sense.

With these statements of scientific methodology in mind, we return to the specific problem of construct validity as applied to psychological tests. The preceding guide rules should reassure the "toughminded," who fear that allowing construct validation opens the door to nonconfirmable test claims. *The answer is that unless the network makes contact with observations, and exhibits explicit, public steps of inference, construct validation cannot be claimed*. An admissible psychological construct must be behavior-relevant (59, p. 15). For most tests intended to measure constructs, adequate criteria do not exist. This being the case, many such tests have been left unvalidated, or a finespun network of rationalizations has been offered as if it were validation. Rationalization is not construct validation. One who claims that his test reflects a construct cannot maintain his claim in the face of recurrent negative results because these results show that his construct is too loosely defined to yield verifiable inferences.

A rigorous (though perhaps probabilistic) chain of inference is required to establish a test as a measure of a construct. To validate a claim that a test measures a construct, a nomological net surrounding the concept must exist. When a construct is fairly new, there may be few specifiable associations by which to pin down the concept. As research proceeds, the construct sends out roots in many directions, which attach it to more and more facts or other constructs. Thus the electron has more accepted properties than the neutrino: *numerical ability* has more than *the second space factor*.

"Acceptance," which was critical in criterion-oriented and content validities, has now appeared in construct validity. Unless substantially the same nomological net is accepted by the several users of the construct, public validation is impossible. If A uses *aggressiveness* to mean overt assault on others, and B's usage includes repressed hostile reactions, evidence

which convinces B that a test measures *aggressiveness* convinces A that the test does not. Hence, the investigator who proposes to establish a test as a measure of a construct must specify his network or theory sufficiently clearly that others can accept or reject it (cf. 41, p. 406). A consumer of the test who rejects the author's theory cannot accept the author's validation. He must validate the test for himself, if he wishes to show that it represents the construct as *he* defines it.

Two general qualifications are in order with reference to the methodological principles 1-6 set forth at the beginning of this section. Both of them concern the amount of "theory," in any high-level sense of that word, which enters into a construct-defining network of laws or lawlike statements. We do not wish [p. 292] to convey the impression that one always has a very elaborate theoretical network, rich in hypothetical processes or entities.

Constructs as inductive summaries. In the early stages of development of a construct or even at more advanced stages when our orientation is thoroughly practical, little or no theory in the usual sense of the word need be involved. In the extreme case the hypothesized laws are formulated entirely in terms of descriptive (observational) dimensions although not all of the relevant observations have actually been made.

The hypothesized network "goes beyond the data" only in the limited sense that it purports to *characterize* the behavior facets which belong to an observable but as yet only partially sampled cluster; hence, it generates predictions about hitherto unsampled regions of the phenotypic space. Even though no unobservables or high-order theoretical constructs are introduced, an element of inductive extrapolation appears in the claim that a cluster including some elements not-yet-observed has been identified. Since, as in any sorting or abstracting task involving a finite set of complex elements, several nonequivalent bases of categorization are available, the investigator may choose a hypothesis which generates erroneous predictions. The failure of a supposed, hitherto untried, member of the cluster to behave in the manner said to be characteristic of the group, or the finding that a nonmember of the postulated cluster does behave in this manner, may modify greatly our tentative construct.

For example, one might build an intelligence test on the basis of his background notions of "intellect," including vocabulary, arithmetic calculation, general information, similarities, two-point threshold, reaction time, and line bisection as subtests. The first four of these correlate, and he extracts a huge first factor. This becomes a second approximation of the intelligence construct, described by its pattern of loadings on the four tests. The other three tests have negligible loading on any common factor. On this evidence the investigator reinterprets intelligence as "manipulation of words." Subsequently it is discovered that test-stupid people are rated as unable to express their ideas, are easily taken in by fallacious arguments, and misread complex directions. These data support the "linguistic" definition of intelligence and the test's claim of validity *for* that construct. But then a block design test with pantomime instructions is found to be strongly saturated with the first factor. Immediately the purely "linguistic" interpretation of Factor I becomes suspect. This finding, taken together with our initial acceptance of the others as relevant to the background concept of intelligence, forces us to reinterpret the concept once again.

If we simply *list* the tests or traits which have been shown to be saturated with the "factor" or which belong to the cluster, no construct is employed. As soon as we even *summarize the properties* of this group of indicators -- we are already making some guesses. Intensional characterization of a domain is hazardous since it selects (abstracts) properties and implies that new tests sharing those properties will behave as do the known tests in the cluster, and that tests not sharing them will not.

The difficulties in merely "characterizing the surface cluster" are strikingly exhibited by the use of certain special and extreme groups for purposes of construct validation. The P_d scale of MMPI was originally de- [p. 293] rived and cross-validated upon hospitalized patients diagnosed "Psychopathic personality, asocial and amoral type" (42). Further research shows the scale to have a limited degree of predictive and concurrent validity for "delinquency" more broadly defined (5, 28). Several studies show associations between P_d and very special "criterion" groups which it would be ludicrous to identify as "*the* criterion" in the traditional sense. If one lists these heterogeneous groups and tries to characterize them intensionally, he faces enormous conceptual difficulties. For example, a recent survey of hunting accidents in Minnesota showed that hunters who had "carelessly" shot someone were significantly elevated on P_d when compared with other hunters (48). This is in line with one's theoretical expectations; when you ask MMPI "experts" to predict for such a group they invariably predict P_d or M_a or both. The finding seems therefore to lend some slight support to the construct validity of the P_d scale. But of course it would be nonsense to *define* the P_d component "operationally" in terms of, say, accident proneness. We might try to subsume the original phenotype and the hunting-accident proneness under some broader category, such as "Disposition to violate society's rules, whether legal, moral, or just *sensible*." But now we have ceased to have a neat operational criterion, and are using instead a rather vague and wide-range class. Besides, there is worse to come. We want the class specification to cover a group trend that (nondelinquent) high school students judged by their peer group as least "responsible" score over a full sigma higher on P_d than those judged most "responsible" (23, p. 75). Most of the behaviors contributing to such sociometric choices fall well within the range of socially permissible action; the proffered criterion specification is still too restrictive. Again, any clinician familiar with MMPI lore would predict an elevated P_d on a sample of (nondelinquent) professional actors. Chyatte's confirmation of this prediction (10) tends to support *both*: (*a*) the theory sketch of "what the P_d factor is, psychologically"; and (*b*) the claim of the P_d scale to construct validity for this hypothetical factor. Let the reader try his hand at writing a brief phenotypic criterion specification that will cover both trigger-happy hunters and Broadway actors! And if he should be ingenious enough to achieve this, does his definition also encompass Hovey's report that high P_d predicts the judgments "not shy" and "unafraid of mental patients" made upon nurses by their supervisors (32, p. 143)? And then we have Gough's report that *low* P_d is associated with ratings as "good-natured" (24, p. 40), and Roessell's data showing that high P_d is predictive of "dropping out of high school" (54). The point is that all seven of these "criterion" dispositions would be readily guessed by any clinician having even superficial familiarity with MMPI interpretation; but to mediate these inferences explicitly requires quite a few hypotheses about dynamics, constituting an admittedly sketchy (but far from vacuous) network defining the genotype *psychopathic deviate*.

Vagueness of present psychological laws. This line of thought leads directly to our second important qualification upon the network schema. The idealized picture is one of a tidy set of postulates which jointly entail the desired theorems; since some of the theorems are coordinated to the observation base, the system constitutes an implicit definition of the [p. 294] theoretical primitives and gives them an indirect empirical meaning. In practice, of course, even the most advanced physical sciences only approximate this ideal. Questions of "categoricalness" and the like, such as logicians raise about pure calculi, are hardly even statable for empirical networks. (What, for example, would be the desiderata of a "well-formed formula" in molar behavior theory?) Psychology works with crude, half-explicit formulations. We do not worry about such advanced formal questions as "whether all molar-behavior statements are decidable by appeal to the postulates" because we know that no existing theoretical network suffices to predict even the *known* descriptive laws. Nevertheless, the sketch of a network is there; if it were not, we would not be saying *anything* intelligible about our constructs. We do not have the rigorous implicit definitions of formal calculi (which still, be it noted, usually permit of a multiplicity of interpretations). Yet the vague, avowedly incomplete network still gives the constructs whatever meaning they do have. When the network is very incomplete, having many strands missing entirely and some constructs tied in only by tenuous threads, then the "implicit definition" of these constructs is disturbingly loose; one might say that the meaning of the constructs is underdetermined. *Since the meaning of theoretical constructs is set forth by stating the laws in which they occur, our incomplete knowledge of the laws of nature produces a vagueness in our constructs* (see Hempel, 30; Kaplan, 34; Pap, 51). We will be able to say "what anxiety is" when we know all of the laws involving it; meanwhile, since we are in the process of discovering these laws, we do not yet know precisely what anxiety is.

CONCLUSIONS REGARDING THE NETWORK AFTER EXPERIMENTATION

The proposition that *x* per cent of test variance is accounted for by the construct is inserted into the accepted network. The network then generates a testable prediction about the relation of the tests scores to certain other variables, and the investigator gathers data. If prediction and result are in harmony, he can retain his belief that the test measures the construct. The construct is at best adopted, never demonstrated to be "correct."

We do not first "prove" the theory, and then validate the test, nor conversely. In any probable inductive type of inference from a pattern of observations, we examine the relation between the total network of theory and observations. The system involves propositions relating test to construct, construct to other constructs, and finally relating some of these constructs to observables. In ongoing research the chain of inference is very complicated. Kelly and Fiske (36, p. 124) give a complex diagram showing the numerous inferences required in validating a prediction from assessment techniques, where theories about the criterion situation are as integral a part of the prediction as are the test data. A predicted empirical relationship permits us to test all the propositions leading to that prediction. Traditionally the proposition claiming to interpret the test has been set apart as the hypothesis being tested, but actually the evidence is significant for all parts of the chain. If the prediction is not confirmed, any link in the chain may be wrong.

A theoretical network can be divided into subtheories used in making particular predictions. All the events successfully predicted through a subtheory are of course evidence in favor of that theory. Such a subtheory [p. 295] may be so well confirmed by voluminous and diverse evidence that we can reasonably view a particular experiment as relevant only to the test's validity. If the theory, combined with a proposed test interpretation, mispredicts in this case, it is the latter which must be abandoned. On the other hand, the accumulated evidence for a test's construct validity may be so strong that an instance of misprediction will force us to modify the subtheory employing the construct rather than deny the claim that the test measures the construct.

Most cases in psychology today lie somewhere between these extremes. Thus, suppose we fail to find a greater incidence of "homosexual signs" in the Rorschach records of paranoid patients. Which is more strongly disconfirmed -- the Rorschach signs or the orthodox theory of paranoia? The negative finding shows the bridge between the two to be undependable, but this is all we can say. The bridge cannot be used unless one end is placed on solider ground. The investigator must decide which end it is best to relocate.

Numerous successful predictions dealing with phenotypically diverse "criteria" give greater weight to the claim of construct validity than do fewer predictions, or predictions involving very similar behaviors. In arriving at diverse predictions, the hypothesis of test validity is connected each time to a subnetwork largely independent of the portion previously used. Success of these derivations testifies to the inductive power of the test-validity statement, and renders it unlikely that an equally effective alternative can be offered.

Implications of Negative Evidence

The investigator whose prediction and data are discordant must make strategic decisions. His result can be interpreted in three ways:

1. The test does not measure the construct variable.

2. The theoretical network which generated the hypothesis is incorrect.

3. The experimental design failed to test the hypothesis properly. (Strictly speaking this may be analyzed as a special case of 2, but in practice the distinction is worth making.)

For further research. If a specific fault of procedure makes the third a reasonable possibility, his proper response is to perform an adequate study, meanwhile making no report. When faced with the other two alternatives, he may decide that his test does not measure the construct adequately. Following that decision, he will perhaps prepare and validate a new test. Any rescoring or new interpretative procedure for the original instrument, like a new test, requires validation *by means of a fresh body of data.*

The investigator may regard interpretation 2 as more likely to lead to eventual advances. It is legitimate for the investigator to call the network defining the construct into question, if he has confidence in the test. Should the investigator decide that some step in the network

is unsound, he may be able to invent an alternative network. Perhaps he modifies the network by splitting a concept into two or more portions, e.g., by designating types of *anxiety*, or perhaps he specifies added conditions under which a generalization holds. When an investigator modifies the theory in such a manner, he is now required to *gather a fresh body of data* to test the altered hypotheses. This step should normally precede publication of the modified theory. If the new data are consistent with the modified network, he is free from the fear that [p. 296] his nomologicals were gerrymandered to fit the peculiarities of his first sample of observations. He can now trust his test to some extent, because his test results behave as predicted.

The choice among alternatives, like any strategic decision, is a gamble as to which course of action is the best investment of effort. Is it wise to modify the theory? That depends on how well the system is confirmed by prior data, and how well the modifications fit available observations. Is it worth while to modify the test in the hope that it will fit the construct? That depends on how much evidence there is -- apart from this abortive experiment -- to support the hope, and also on how much it is worth to the investigator's ego to salvage the test. The choice among alternatives is a matter of research planning.

For practical use of the test. The consumer can accept a test as a measure of a construct only when there is a strong positive fit between predictions and subsequent data. When the evidence from a proper investigation of a published test is essentially negative, it should be reported as a stop sign to discourage use of the test pending a reconciliation of test and construct, or final abandonment of the test. If the test has not been published, it should be restricted to research use until some degree of validity is established (1). The consumer can await the results of the investigator's gamble with confidence that proper application of the scientific method will ultimately tell whether the test has value. Until the evidence is in, he has no justification for employing the test as a basis for terminal decisions. The test may serve, at best, only as a source of suggestions about individuals to be confirmed by other evidence (15, 47).

There are two perspectives in test validation. From the viewpoint of the psychological practitioner, the burden of proof is on the test. A test should not be used to measure a trait until its proponent establishes that predictions made from such measures are consistent with the best available theory of the trait. In the view of the test developer, however, both the test and the theory are under scrutiny. He is free to say *to himself privately*, "If my test disagrees with the theory, so much the worse for the theory." This way lies delusion, unless he continues his research using a better theory.

Reporting of Positive Results

The test developer who finds positive correspondence between his proposed interpretation and data is expected to report the basis for his validity claim. Defending a claim of construct validity is a major task, not to be satisfied by a discourse without data. The *Technical Recommendations* have little to say on reporting of construct validity. Indeed, the only detailed suggestions under that heading refer to correlations of the test with other measures, together with a cross reference to some other sections of the report. The two key principles, however, call for the most comprehensive type of reporting. The manual for any

test "should report all available information which will assist the user in determining what psychological attributes account for variance in test scores" (59, p. 27). And, "The manual for a test which is used primarily to assess postulated attributes of the individual should outline the theory on which the test is based and organize whatever partial validity data there are to show in what way they support the theory" (59, p. 28). It is recognized, by a classification as "very desirable" rather than "essential," that in the latter recom- [p. 297] mendation goes beyond present practice of test authors.

The proper goals in reporting construct validation are to make clear (*a*) what interpretation is proposed, (*b*) how adequately the writer believes this interpretation is substantiated, and (*c*) what evidence and reasoning lead him to this belief. Without *a* the construct validity of the test is of no use to the consumer. Without *b* the consumer must carry the entire burden of evaluating the test research. Without *c* the consumer or reviewer is being asked to take *a* and *b* on faith. The test manual cannot always present an exhaustive statement on these points, but it should summarize and indicate where complete statements may be found.

To specify the interpretation, the writer must state what construct he has in mind, and what meaning he gives to that construct. For a construct which has a short history and has built up few connotations, it will be fairly easy to indicate the presumed properties of the construct, i.e., the nomologicals in which it appears. For a construct with a longer history, a summary of properties and references to previous theoretical discussions may be appropriate. It is especially critical to distinguish proposed interpretations from other meanings previously given the same construct. The validator faces no small task; he must somehow communicate a theory to his reader.

To evaluate his evidence calls for a statement like the conclusions from a program of research, noting what is well substantiated and what alternative interpretations have been considered and rejected. The writer must note what portions of his proposed interpretation are speculations, extrapolations, or conclusions from insufficient data. The author has an ethical responsibility to prevent unsubstantiated interpretations from appearing as truths. A claim is unsubstantiated unless the evidence for the claim is public, so that other scientists may review the evidence, criticize the conclusions, and offer alternative interpretations.

The report of evidence in a test manual must be as complete as any research report, except where adequate public reports can be cited. Reference to something "observed by the writer in many clinical cases" is worthless as evidence. Full case reports, on the other hand, may be a valuable source of evidence so long as these cases are representative and negative instances receive due attention. The report of evidence must be interpreted with reference to the theoretical network in such a manner that the reader sees why the author regards a particular correlation or experiment as confirming (or throwing doubt upon) the proposed interpretation. Evidence collected by others must be taken fairly into account.

VALIDATION OF A COMPLEX TEST "AS A WHOLE"

Special questions must be considered when we are investigating the validity of a test which is aimed to provide information about several constructs. In one sense, it is naive to inquire "Is this test valid?" One does not validate a test, but only a principle for making inferences.

If a test yields many different types of inferences, some of them can be valid and others invalid (cf. Technical Recommendation C2: "The manual should report the validity of each type of inference for which a test is recommended"). From this point of view, every topic sentence in the typical book on Rorschach interpretation presents a hypothesis re- [p. 298] quiring validation, and one should validate inferences about each aspect of the personality separately and in turn, just as he would want information on the validity (concurrent or predictive) for each scale of MMPI.

There is, however, another defensible point of view. If a test is purely empirical, based strictly on observed connections between response to an item and some criterion, then of course the validity of one scoring key for the test does not make validation for its other scoring keys any less necessary. But a test may be developed on the basis of a theory which in itself provides a linkage between the various keys and the various criteria. Thus, while Strong's Vocational Interest Blank is developed empirically, it also rests on a "theory" that a youth can be expected to be satisfied in an occupation if he has interests common to men now happy in the occupation. When Strong finds that those with high Engineering interests scores in college are preponderantly in engineering careers 19 years later, he has partly validated the proposed use of the Engineer score (predictive validity). Since the evidence is consistent with the theory on which all the test keys were built, this evidence alone increases the presumption that the *other* keys have predictive validity. How strong is this presumption? Not very, from the viewpoint of the traditional skepticism of science. Engineering interests may stabilize early, while interests in art or management or social work are still unstable. A claim cannot be made that the whole Strong approach is valid just because one score shows predictive validity. But if thirty interest scores were investigated longitudinally and all of them showed the type of validity predicted by Strong's theory, we would indeed by caviling to say that this evidence gives no confidence in the long-range validity of the thirty-first score.

Confidence in a theory is increased as more relevant evidence confirms it, but it is always possible that tomorrow's investigation will render the theory obsolete. The Technical Recommendations suggest a rule of reason, and ask for evidence for each *type* of inference for each *type* of inference for which a test is recommended. It is stated that no test developer can present predictive validities for all possible criteria; similarly, no developer can run all possible experimental tests of his proposed interpretation. But the recommendation is more subtle than advice that a lot of validation is better than a little.

Consider the Rorschach test. It is used for many inferences, made by means of nomological networks at several levels. At a low level are the simple unrationalized correspondences presumed to exist between certain signs and psychiatric diagnoses. Validating such a sign does nothing to substantiate Rorschach theory. For other Rorschach formulas an explicit a priori rationale exists (for instance, high *F*% interpreted as implying rigid control of impulses). Each time such a sign shows correspondence with criteria, its rationale is supported just a little. At a still higher level of abstraction, a considerable body of theory surrounds the general area of *outer control*, interlacing many different constructs. As evidence cumulates, one should be able to decide what specific inference-making chains within this system can be depended upon. One should also be able to conclude -- or deny --

that so much of the system has stood up under test that one has some confidence in even the untested lines in the network.

In addition to relatively delimited nomological networks surrounding [p. 299] *control* or *aspiration*, the Rorschach interpreter usually has an overriding theory of the test as a whole. This may be a psychoanalytic theory, a theory of perception and set, or a theory stated in terms of learned habit patterns. Whatever the theory of the interpreter, whenever he validates an inference from the system, he obtains some reason for added confidence in his overriding system. His total theory is not tested, however, by experiments dealing with only one limited set of constructs. The test developer must investigate far-separated, independent sections of the network. The more diversified the predictions the system is required to make, the greater confidence we can have that only minor parts of the system will later prove faulty. Here we begin to glimpse a logic to defend the judgment that the test and its whole interpretative system is valid at some level of confidence.

There are enthusiasts who would conclude from the foregoing paragraphs that since there is some evidence of correct, diverse predictions made from the Rorschach, the test as a whole can be accepted as validated. This conclusion overlooks the negative evidence. Just one finding contrary to expectation, based on sound research, is sufficient to wash a whole theoretical structure away. Perhaps the remains can be salvaged to form a new structure. But this structure now must be exposed to fresh risks, and sound negative evidence will destroy it in turn. There is sufficient negative evidence to prevent acceptance of the Rorschach and its accompanying interpretative structures as a whole. So long as any aspects of the overriding theory stated for the test have been disconfirmed, this structure must be rebuilt.

Talk of areas and structures may seem not to recognize those who would interpret the personality "globally." They may argue that a test is best validated in matching studies. Without going into detailed questions of matching methodology, we can ask whether such a study validates the nomological network "as a whole." The judge does employ some network in arriving at his conception of his subject, integrating specific inferences from specific data. Matching studies, if successful, demonstrate only that each judge's interpretative theory has some validity, that it is not completely a fantasy. Very high consistency between judges is required to show that they are using the same network, and very high success in matching is required to show that the network is dependable.

If inference is less than perfectly dependable, we must know which aspects of the interpretative network are least dependable and which are most dependable. Thus, even if one has considerable confidence in a test "as a whole" because of frequent successful inferences, one still returns as an ultimate aim to the request of the Technical Recommendation for separate evidence on the validity of each type of inference to be made.

RECAPITULATION

Construct validation was introduced in order to specify types of research required in developing tests for which the conventional views on validation are inappropriate.

Personality tests, and some tests of ability, are interpreted in terms of attributes for which there is no adequate criterion. This paper indicates what sorts of evidence can substantiate such an interpretation, and how such evidence is to be interpreted. The following points made in the discussion are particularly significant.

1. A construct is defined implicitly by a network of associations or propo- [p. 300] sitions in which it occurs. Constructs employed at different stages of research vary in definiteness.

2. Construct validation is possible only when some of the statements in the network lead to predicted relations among observables. While some observables may be regarded as "criteria," the construct validity of the criteria themselves is regarded as under investigation.

3. The network defining the construct, and the derivation leading to the predicted observation, must be reasonably explicit so that validating evidence may be properly interpreted.

4. Many types of evidence are relevant to construct validity, including content validity, interitem correlations, intertest correlations, test-"criterion" correlations, studies of stability over time, and stability under experimental intervention. High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct.

5. When a predicted relation fails to occur, the fault may lie in the proposed interpretation of the test or in the network. Altering the network so that it can cope with the new observations is, in effect, redefining the construct. Any such new interpretation of the test must be validated by a fresh body of data before being advanced publicly. Great care is required to avoid substituting a posteriori rationalizations for proper validation.

6. Construct validity cannot generally be expressed in the form of a single simple coefficient. The data often permit one to establish upper and lower bounds for the proportion of test variance which can be attributed to the construct. The integration of diverse data into a proper interpretation cannot be an entirely quantitative process.

7. Constructs may vary in nature from those very close to "pure description" (involving little more than extrapolation of relations among observation-variables) to highly theoretical constructs involving hypothesized entities and processes, or making identifications with constructs of other sciences.

8. The investigation of a test's construct validity is not essentially different from the general scientific procedures for developing and confirming theories.

Without in the least *advocating* construct validity as preferable to the other three kinds (concurrent, predictive, content), we do believe it imperative that psychologists make a place for it in their methodological thinking, so that its rationale, its scientific legitimacy, and its dangers may become explicit and familiar. This would be preferable to the

widespread current tendency to engage in what actually amounts to construct validation research and use of constructs in practical testing, while talking an "operational" methodology which, if adopted, would force research into a mold it does not fit.

Footnotes

[1] The second author worked on this problem in connection with his appointment to the Minnesota Center for Philosophy of Science. We are indebted to the other members of the Center (Herbert Feigl, Michael Scriven, Wilfrid Sellars), and to D. L. Thistlethwaite of the University of Illinois, for their major contributions to our thinking and their suggestions for improving this paper.

[2] Referred to in a preliminary report (58) as *congruent validity.*

REFERENCES

1. AMERICAN PSYCHOLOGICAL ASSOCIATION. *Ethical standards of psychologists.* Washington, D.C.: American Psychological Association, Inc., 1953.

2. ANASTASI, ANNE. The concept of validity in the interpretation of test scores. *Educ. psychol. Measmt*, 1950, 10, 67-78.

3. BECHTOLDT, H. P. Selection. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 1237-1267.

4. BECK, L. W. Constructions and inferred entities. *Phil. Sci*., 1950, 17. Reprinted in H. Feigl and M. Brodbeck (Eds.), *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 368-381.

5. BLAIR, W. R. N. A comparative study of disciplinary offenders and non-offenders in the Canadian Army. *Canad. J. Psychol.*, 1950, 4, 49-62.

6. BRAITHWAITE, R. B. *Scientific explanation*. Cambridge: Cambridge Univer. Press, 1953.

7. CARNAP, R. Empiricism, semantics, and ontology. *Rév. int. de Phil*., 1950, II, 20-40. Reprinted in P. P. Wiener (Ed.), *Readings in philosophy of science*, New York: Scribner's, 1953. Pp. 509-521.

8. CARNAP, R. *Foundations of logic and mathematics*. *International encyclopedia of unified science*, I, No. 3. Pages 56-69 reprinted as "The interpretation of physics" in H.

Feigl and M. Brodbeck (Eds.), *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 309-318.

9. CHILD, I. L. Personality. *Annu. Rev. Psychol*., 1954, 5, 149-171.

10. CHYATTE, C. Psychological characteristics of a group of professional actors. *Occupations*, 1949, 27, 245-250.

11. CRONBACH, L. J. *Essentials of psychological testing*. New York: Harper, 1949.

12. CRONBACH, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt*, 1950, 10, 3-31.

13. CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-335.

14. CRONBACH, L. J. Processes affecting scores on "understanding of others" and "assumed similarity." *Psychol. Bull*., 1955, 52, 177-193.

15. CRONBACH, L. J. The counselor's problems from the perspective of communication theory. In Vivian H. Hewer (Ed.), *New perspectives in counseling*. Minneapolis: Univer. of Minnesota Press, 1955.

16. CURETON, E. E. Validity. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1950. Pp. 621-695.

17. DAMRIN, DORA E. A comparative study of information derived from a diagnostic problem-solving test by logical and factorial methods of scoring. Unpublished doctor's dissertation, Univer. of Illinois, 1952.

18. EYSENCK, H. J. Criterion analysis -- an application of the hypothetico-deductive method in factor analysis. *Psychol. Rev*., 1950, 57, 38-53.

19. FEIGL, H. Existential hypotheses. *Phil. Sci*., 1950, 17, 35-62.

20. FEIGL, H. Confirmability and confirmation. *Rév. int. de Phil*., 1951, 5, 1-12. Reprinted in P. P. Wiener (Ed.), *Readings in philosophy of science*. New York: Scribner's, 1953. Pp. 522-530.

21. GAYLORD, R. H. Conceptual consistency and criterion equivalence: a dual approach to criterion analysis. Unpublished manuscript (PRB Research Note No. 17). Copies obtainable from ASTIA-DSC, AD-21 440.

22. GOODENOUGH, FLORENCE L. *Mental testing*. New York: Rinehart, 1950.

23. GOUGH, H. G., McCLOSKY, H., & MEEHL, P. E. A personality scale for social responsibility. *J. abnorm. soc. Psychol.*, 1952, 47, 73-80.

24. GOUGH, H. G., McKEE, M. G., & YANDELL, R. J. Adjective check list analyses of a number of selected psychometric and assessment variables. Unpublished manuscript. Berkeley: IPAR, 1953.

25. GUILFORD, J. P. New standards for test evaluation. *Educ. psychol. Measmt*. 1946, 6, 427-439.

26. GUILFORD, J. P. Factor analysis in a test-development program. *Psychol. Rev.*, 1948, 55, 79-94.

27. GULLIKSEN, H. Intrinsic validity. *Amer. Psychologist*, 1950, 5, 511-517.

28. HATHAWAY, S. R., & MONACHESI, E. D. *Analyzing and predicting juvenile delinquency with the MMPI.* Minneapolis: Univer. of Minnesota Press, 1953.

29. HEMPEL, C. G. Problems and changes in the empiricist criterion of meaning. *Rév. int. de Phil*., 1950, 4, 41-63. Reprinted in L. Linsky, *Semantics and the philosophy of language*. Urbana: Univer. of Illinois Press, 1952. Pp. 163-185.

30. HEMPEL, C. G. *Fundamental of concept formation in empirical science*. Chicago: Univer. of Chicago Press, 1952.

31. HORST, P. The prediction of personal adjustment. *Soc. Sci. Res. Council Bull*., 1941, No. 48.

32. HOVEY, H. B. MMPI profiles and personality characteristics. *J. consult. Psychol*., 1953, 17, 142-146.

33. JENKINS, J. G. Validity for what? *J. consult. Psychol*., 1946, 10, 93-98.

34. KAPLAN, A. Definition and specification of meaning. *J. Phil*., 1946, 43, 281-288.

35. KELLY, E. L. Theory and techniques of assessment. *Annu. Rev. Psychol*., 1954, 5, 281-311.

36. KELLY, E. L., & FISKE, D. W. *The prediction of performance in clinical psychology*. Ann Arbor: Univer. of Michigan Press, 1951.

37. KNEALE, W. *Probability and induction*. Oxford: Clarendon Press, 1949. Pages 92-110 reprinted as "Induction, explanation, and transcendent hypotheses" in H. Feigl and M. Brodbeck (Eds.), *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 353-367.

38. LINDQUIST, E. F. *Educational measurement*. Washington, D.C.: American Council on Education, 1950.

39. LUCAS, C. M. Analysis of the relative movement test by a method of individual interviews. *Bur. Naval Personnel Res. Rep*., Contract Nonr-694 (00), NR 151-13, Educational Testing Service, March 1953.

40. MacCORQUODALE, K., & MEEHL, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev*., 1948, 55, 95-107.

41. MacFARLANE, JEAN W. Problems of validation inherent in projective methods. *Amer. J. Orthopsychiat*., 1942, 12, 405-410.

42. McKINLEY, J. C., & HATHAWAY, S. R. The MMPI: V. Hysteria, hypomania, and psychopathic deviate. *J. appl. Psychol*., 1944, 28, 153-174.

43. McKINLEY, J. C., HATHAWAY, S. R., & MEEHL, P. E. The MMPI: VI. The K scale. *J. consult. Psychol*., 1948, 12, 20-31.

44. MEEHL, P. E. A simple algebraic development of Horst's suppressor variables. *Amer. J. Psychol*., 1945, 58, 550-554.

45. MEEHL, P. E. An investigation of a general normality or control factor in personality testing. *Psychol. Monogr*., 1945, 59, No. 4 (Whole No. 274).

46. MEEHL, P. E. *Clinical vs. statistical prediction*. Minneapolis: Univer. of Minnesota Press, 1954.

47. MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychol. Bull*., 1955, 52, 194-216.

48. *Minnesota Hunter Casualty Study*. St. Paul: Jacob Schmidt Brewing Company, 1954.

49. MOSIER, C. I. A critical examination of the concepts of face validity. *Educ. psychol. Measmt*, 1947, 7, 191-205.

50. MOSIER, C. I. Problems and designs of cross-validation. *Educ. psychol. Measmt*, 1951, 11, 5-12.

51. PAP, A. Reduction-sentences and open concepts. *Methodos*, 1953, 5, 3-30.

52. PEAK HELEN. Problems of objective observation. In L. Festinger and D. Katz (Eds.), *Research methods in the behavioral sciences*. New York: Dryden Press, 1953, Pp. 243-300.

53. PORTEUS, S. D. *The Porteus maze test and intelligence*. Palo Alto: Pacific Books, 1950.

54. ROESSEL, F. P. MMPI results for high school drop-outs and graduates. Unpublished doctor's dissertation, Univer. of Minnesota, 1954.

55. SELLARS, W. S. Concepts as involving laws and inconceivable without them. *Phil. Sci.*, 1948, 15, 287-315.

56. SELLARS, W. S. Some reflections on language games. *Phil. Sci.*, 1954, 21, 204-228.

57. SPIKER, C. C., & McCANDLESS, B. R. The concept of intelligence and the philosophy of science. *Psychol. Rev.*, 1954, 61, 255-267.

58. Technical recommendations for psychological tests and diagnostic techniques: preliminary proposal. *Amer. Psychologist*, 1952, 7, 461-476.

59. Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull. Supplement*, 1954, 51, 2, Part 2, 1-38.

60. THURSTONE, L. L. The criterion problem in personality research. *Psychometric Lab. Rep.*, No. 78. Chicago: Univer. of Chicago, 1952.

Received for early publication February 18, 1955.