Chapter 5

Objective Tests as Instruments of Psychological Theory

JANE LOEVINGER

The central concepts of classical test theory are reliability and validity. The concept of reliability was criticized from the beginning, and at present the most widely accepted view appears to be that what was formerly called reliability encompasses two independent components, homogeneity and stability. The notion of validity as correlation with an outside criterion appeared to be a simpler and less vulnerable concept than reliability, however difficult it was to obtain adequate criteria. The rather sudden appearance of the term "construct validity" indicates that the second concept of classical test theory is undergoing criticism and revision.

The purposes of the present monograph are:

(a) to celebrate the extension of the concept of validity as an indication that psychometrics is recognized as truly the handmaiden of psychology rather than merely of psychotechnology;

(b) to argue that, since predictive, concurrent, and content validities are all essentially *ad hoc*, construct validity is the whole of validity from a scientific point of view;

(c) to analyze the components of construct validity, in particular propos-

This work is reprinted from *Psychological Reports*, 1957, Monograph Supplement 9, with the permission of the author and Southern University Press.

This monograph was written in part while the author held the Margaret M. Justin Fellowship of the American Association of University Women. This investigation was also supported in part by a research grant, M-1213, from the National Institute of Mental Health, of the National Institutes of Health, Public Health Service.

The author owes more than an ordinary debt of gratitude for detailed criticism of early versions of the manuscript to Drs. Jack Block, Clyde Coombs, Lee Cronbach, Louis Guttman, Paul Meehl, Blanche Sweet, and Robert M. W. Travers. They have helped to increase areas of mutual agreement and sharpen points of disagreement. Permission to quote unpublished material has been granted by Drs. Block, Sweet, and Milton Whitcomb.

While this monograph was developed independently, it is a pleasure to record the priority of Jessor and Hammond [60] with respect to some of the ideas presented here, particularly that the concept of construct validity implies a program not only of test analysis but also of test construction.

ing "structural component" as name for a previously only partly recognized aspect; and

(d) to relate secular trends in test behavior to the validity problem.

The presentation of these and related topics as a single monograph has an overriding purpose: to develop a coherent view of psychometrics, a mutually implicative test theory and method of test construction. Many current lines of research in psychometrics are discussed, either as contributing to or as contrasting with the present system. It would be unjust, however, to read the monograph as a review of psychometric literature. No attempt is made to evaluate any contribution or any line of work *in toto*, nor is inclusive coverage of current literature sought. If the exposition appears to claim that only the present view is admissible, that is an artifact of the argument, not an expression of belief or intention.

EXTENSION OF THE CONCEPT OF VALIDITY

Because of the difficulties to which the classical concept of validity led, there have been many attempts to modify and redefine validity in recent years, culminating in recognition that the new concept is essentially different from classical validity and requires an identifying name.

The term *construct validity* was proposed by the APA Committee on Psychological Tests, which drew up the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* [121]. The concept was suggested by a subcommittee composed of Meehl and Challman. It was expounded later by Cronbach and Meehl [20] in a paper which reviewed more fully the history of the concept. The circumstances which gave rise to these two contributions were somewhat restrictive. Obviously, the official *Technical Recommendations* were constrained to approve the best in current practice, and the paper by Cronbach and Meehl grew out of that endeavor. The present monograph, being without official sanction or origin, is in position to take advantage of the work done by official committees and by Cronbach and Meehl but also to propose a more radical reformulation of the validity problem. This reformulation is not intended chiefly as criticism of the previous excellent contributions. The organization of the present monograph follows approximately that of Peak's chapter on objective tests [92], which Cronbach and Meehl also acknowledge as their closest predecessor. The province is large; specific overlap with Peak and with Cronbach and Meehl will therefore be avoided.

A. CRITIQUE OF CLASSICAL VALIDITY CONCEPT

Validity has often been defined as the extent to which a test measures what it is supposed to measure. This definition is too vague, too remote from actual measuring operations, to be useful; it is consistent with all the current meanings of validity. What will be referred to as the classical definition of validity is the one which by far predominates in psychometric literature, to wit, correlation with a criterion.

What the *Technical Recommendations* [121] call predictive and concurrent validities seems indistinguishable from what Peak [92] calls "blindly empirical validity" and is exactly the classical conception of validity in test theory. Surely no one will dispute the legitimacy of computing a validity coefficient for an existing test in reference to a situation where it is deemed a suitable predictive or discriminative device. The contention of the present monograph is that classical validity is not a suitable basic concept for test theory; it does not provide an adequate basis for test construction.

The classical definition of validity has been stated:

> The validity of a test is the correlation of the test with some criterion. In this

> sense a test has a great many different "validities." For example, the ACE Psychological Examination has one validity for predicting grades in English and a different validity for predicting grades in Latin. It is also found in studying various validity coefficients for a given test that they vary from school to school, and from time to time. In other words, validity cannot be regarded as a fixed or a unitary characteristic of a test. As new uses for a test are contemplated, new validity coefficients must be determined; and, when use of a test is continued, the validity coefficients must be redetermined at intervals [43, p. 88].

The above quotation is the essence of the single paragraph devoted to the "Meaning of validity" in Gulliksen's text on test theory.

Meehl and Rosen [89], in a review of current uses of predictive and concurrent validity coefficients, have carried further the demonstration of the specificity of classical validity. They are particularly concerned with the case where the criterion to be predicted is a dichotomous one. In such prediction, or discrimination, there are two possible errors: one can select an individual who should not be selected or reject an individual who should not be rejected, meaning by selection and rejection no more than assignment to the two distinguished groups.

The dichotomous criterion is particularly important for test theory, for one of the commonest methods of test construction is to select those items which best distinguish two groups believed to differ with respect to the dimension to be measured. In this kind of empirical keying, the ability of an item to discriminate the two criterion groups is the chief or only basis for its inclusion in the test.

Three different cases of dichotomous criteria must be distinguished. A truly dichotomous criterion is one which is not reasonably conceived as two extremes of a continuum. Examples are not easy to find. One possibility is a series of patients all of whom complain of headaches; the problem is to separate those whose complaint is at least partly organic from those whose symptoms are entirely psychogenic. Again, of all patients discharged from a mental hospital, those readmitted and those not readmitted are substantially distinct groups.

A second case might be referred to as a dichotomized criterion. Here the problem is to predict which individuals fall above a given cutting score in an essentially continuous criterion. In studies done in a military setting, passing or failing a given course of study is often the criterion to be predicted. Presumably the course grades initially fall into a more or less continuous distribution. While passing or failing a course is in a way similar to being readmitted to a mental hospital or not, the cutting point between "pass" and "fail" seems less arbitrary, less subject to administrative whim, in the case of hospital readmission.

In a third case, and probably the most common one, an essentially continuous distribution is being measured and the test is expected ultimately to discriminate throughout its range. For item selection, however, only individuals at the two extremes of the distribution are used.

Consider the first case, a truly dichotomous criterion. Meehl and Rosen [89] have shown that validity measured in terms of false positives and false negatives is altered when nothing changes except the proportion of true positives and true negatives. The optimal cutting score depends on the proportions in the two groups, and for this reason they advocate that an inflexible cutting score not be set for any psychometric device. Merely changing the proportion in the two groups without any alteration in the nature of the individuals included should not affect item choice, however, when items are chosen according to their ability to discriminate the groups.

Consider now the second case, the dichotomized criterion. Any change in administrative conditions which leads to shifting the cutting point on the criterion will not only alter the validity of the total test and shift the optimal cutting score on

the test; it will also change the composition of the criterion groups. Thus, discriminative powers of the items will be differentially affected. In general, some items will become more valid and some less. A test whose items are selected to be those with optimal validity one year will not have optimally valid items in later years if the cut between pass and fail has changed. Lord [76] and Cronbach and Warrington [21] have contributed papers on one relevant point, the dependence of the optimal item difficulty on the criterion cutting point.

In the third case, where only the extremes of the distribution are used to establish the validity of the items, considerations similar to those of the second case prevail. It has been demonstrated in a number of papers [71] that for the situation most usual in psychological testing, i.e., low item intercorrelations, optimal item difficulty is approximately at the median of the distribution. Where only the extremes of the distribution are used to test item validity, insufficient evidence in regard to item difficulty is obtained. It should be noted, too, that apart from the problem of item difficulty, knowledge of the ability of the item to discriminate the extremes of a distribution is not substitutable for knowledge of how accurately it discriminates near the median.

Now a serious test construction project is an arduous and expensive business. No one recommends attempting to construct and standardize tests except on the basis of large samples representative of some specifiable population. Major test construction is rarely if ever undertaken with the intention of putting a test to a single use.

Meehl and Rosen [89] show, however, that even when the administrative or clinical setting remains unaltered in other respects, if the test is used in a slightly different manner, the validity is altered. They give as an example the use of a neuropsychiatric screening test by the Army. If a test is constructed to predict which inductees will later be given neuropsychiatric discharges, it is not thereby validated for selecting inductees. A test constructed to select those draftees for whom examination by a psychiatrist is desirable prior to induction is not thereby validated as a selection instrument for the draftees for whom psychiatrists have difficulty in making a decision to induct into the Army. Such changes in test utilization involve changes in the composition of criterion groups, even though the trait to be measured and the criterion cutting point remain the same. Not only the validity coefficient and the optimal cutting score but also item validities are affected by such changes. A scale selected to discriminate hysterics from normals would be quite different from a scale selected to discriminate hysterics from early schizophrenics.

In short, it is difficult to discover any circumstances under which the classical concept of validity is a suitable basis for test construction. Military situations where large numbers of men are processed in short periods of time offer the most promising possibility of an administrative setting stable enough to justify test construction on the basis of classical validity. But it is well known that administrative fiat can change radically a selection per cent without advance warning. Indeed, apart from any such argument as the foregoing, many tests devised in the military situation have been declared obsolete before they were completed.

Since expositions of test theory such as Gulliksen's [43] *Theory of Mental Tests* devote far more space to reliability than to validity, it may be charged that arguing against classical validity as the basic concept of test theory is to attack a straw man. But reliability as a central concept also leads to problems and contradictions, some of which the writer has called attention to previously [69, 71] and some of which will be discussed under Secular Trends in Tests Behavior, below. In consideration of such problems Brogden [5], among others, argued that validity rather than reliability is the central concept of test theory. In this respect it is interesting to contrast Gul-

liksen's text with the *Technical Recommendations;* where the former is devoted predominantly to reliability theory with validity given only minor emphasis, the more recent official publication devotes about three times the space to validity that it does to reliability.

Lord strongly defends the importance of specific validity, stating that the concept of over-all validity "is not basic to psychometrics. The discriminating power of the test for a specified decision problem regarding a specified examinee is the truly basic concept" [78, p. 509]. Lord's view will be seen to represent the opposite pole from that represented by the present monograph.

A strong proponent of the present view is Cattell, who has written:

> Particularized validation is not only devoid of proper scientific interest but deceptive in its promise of practical economy. . . . Its absurdity is most cogently argued by the demands of practical economy and efficiency alone; for a specific test for every occupation and life situation is its logical and impossible conclusion [11, pp. 549–550].

The writer believes that the most fruitful direction for the development of psychometric devices, and hence of psychometric theory, is toward measurement of traits which have real existence in some sense; that this orientation is antithetical to one which places first emphasis on prediction, decisions, or "utility;" that most decision-oriented psychometric studies would be more fruitfully formulated as trait-oriented studies; and that such legitimately pressing decisions as must inevitably be made will also best be served by a predominantly trait-oriented psychometrics.

An economist, Jacob Marschak, has stated the argument concisely:

> Theory provides us with solutions which are potentially useful for a large class of decisions. It is welcome because we cannot foresee which particular decisions we shall have to take. Our decisions may or may not be such as to leave certain properties of the system unchanged. Hence, the more we know about its properties the better. If we merely want to know how long it takes to boil an egg, the best is to boil one or two without going into the chemistry of protein molecules. The need for chemistry is due to our want to do other and new things [82, p. 214].

The argument against classical criterion-oriented psychometrics is thus two-fold: it contributes no more to the science of psychology than rules for boiling an egg contribute to the science of chemistry. And the number of genuine egg-boiling decisions which clinicians and psychotechnologists face is small compared with the number of situations where a deeper knowledge of psychological theory would be helpful. This argument challenges Meehl's [88] plea for a good clinical cookbook and Cronbach's [18] advocacy of decision and utility theory.

B. CONSTRUCT VALIDITY: ELUCIDATION OF TERMS

As originally proposed, construct validity was one of four kinds of validity, the other three being content, predictive, and concurrent validities. Predictive and concurrent validities are, following the above argument, *ad hoc*. Content validity is established by the judgment of the investigator that the items are valid; it is thus also contingent upon a special, non-generalizable circumstance, to wit, the particular investigator. (But see the fuller discussion of content validity under Components of Construct Validity, A, below.) Since *ad hoc* arguments are scientifically of minor importance, if not actually inadmissable, what is left, construct validity, is the whole of the subject from a systematic, scientific point of view.

Thus, in place of the classification of validity proposed in the *Technical Recommendations,* it is here recommended that two basic contexts for defining validity be recognized, administrative and scientific. There are essentially two kinds of adminis-

trative validity, content and predictive-concurrent. There is only one kind of validity which exhibits the property of transposability or invariance under changes in administrative setting which is the touchstone of scientific usefulness: that is construct validity.

Neither the *Technical Recommendations* nor Cronbach and Meehl gave a formal definition of construct validity. In the former paper the term was introduced as follows: "Construct validity is evaluated by investigating what psychological qualities a test measures, i.e., by demonstrating that certain explanatory constructs account to some degree for performance on the test. . . . Essentially, in studies of construct validity we are validating the theory underlying the test" [121, p. 14].

Cronbach and Meehl's introduction of the term was: "*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?'" [20, p. 282].

The proponents of the term *construct validity* have, I believe, been misled by their philosophical sophistication into using a term less precise and less intuitively appealing than a naively realistic term would be, such as, perhaps, *essential validity*. There are indications of reification of *constructs* among some psychologists to mean general or central traits. Among psychologists who do not like constructs, they apparently stand for a non-preferred level of generalization. The term *trait* seems also to have acquired definite, albeit private, connotations for many psychologists. A useful distinction which MacCorquodale and Meehl [81] have made between intervening variables and hypothetical constructs, in terms of degree of abstraction from the data, was perhaps the jumping off point; but assigning to constructs a particular place in personality organization is a vast extrapolation from their thesis.

A dictionary [122] definition of construct is: "Something constructed; specif., *Psychol.*, an intellectual synthesis." In the present paper both *construct* and *trait* are used in their general or dictionary meanings. Connotations of depth, level, or locus are specifically disclaimed. Traits exist in people; constructs (here usually about traits) exist in the minds and magazines of psychologists. People have constructs too, but that is outside the present scope.

Construct connotes construction and artifice; yet what is at issue is validity with respect to exactly what the psychologist does not construct: the validity of the test as a measure of traits which exist prior to and independently of the psychologist's act of measuring. It is true that psychologists never know traits directly but only through the glass of their constructs, but the data to be judged are manifestations of traits, not manifestations of constructs. Cronbach and Meehl and their colleagues on the APA committee appear reluctant to assign reality to constructs or traits. Considering traits as real is, in the present view, a working stance and not a philosophical tenet.

That the distinction made here between traits and constructs is free of metaphysical implications is seen by comparing it to the familiar distinction between parameter and statistic. The parameter is what we aim to estimate; the corresponding statistic represents our current best estimate of it. Just so, the trait is what we aim to understand, and the corresponding construct represents our current best understanding of it. The distinction between trait and construct can be dispensed with no better than the distinction between parameter and statistic.

Thus, in conceptualizing the problem of validity I find myself far from the naively operational but actually *ad hoc* reasoning of traditional discussions of validity, but objecting to the confusion of constructs and traits which may be read into the term construct validity. With this understanding, the subject of the present monograph is construct validity.

Three elements are involved, the test, the traits measured, and what the tester says

the test measures (construct, interpretation, theory). With three elements, two independent relationships can be specified. This can be done two quite different ways. The *Technical Recommendations* [121, p. 15] imply that the two questions subsumed under construct validity are: to what extent does the test measure whatever is specified by a given construct? And, to what extent does that construct embody a valid hypothesis? An alternative formulation of the two relationships is: to what extent does the test measure a trait that "really" exists? And, how well does the proposed interpretation correspond to what is measured by the test?

The former pair of questions divides construct validity into the validity of the test for the construct and the validity of the construct. Cronbach and Meehl, following Peak, state that what is at stake is usually both, that the evidence for the two kinds of validity is usually not separable, except where either the test or the construct has been established over a long period of time.

The latter pair of questions divides the topic into the intrinsic validity of the test and the validity of the interpretation, though for practical purposes validity usually would also cover validity of the interpretation. The superiority of the latter formulation lies in the fact that information under the two headings comes somewhat separately. (This discussion is not intended to imply that the actual process of test construction should follow such a separation.) The magnitude of the interrelationships of the items among themselves and the magnitude of the highest external correlations of the test score are evidence that something systematic is being measured. The content of the items, the nature of their interrelationships, and the nature of the outside variables showing varying degrees of interrelationship with the test score are relevant to proposed interpretations. The former set of relations may be thought of as giving psychometric meaning to the test, the latter as giving its psychological meaning.

Cronbach and Meehl are led by their adherence to the former pair of questions to state, "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" [20, p. 291]. Do they mean that validational studies are communicable only among such coteries as are agreed on theoretical issues? Such a stand assigns a very minor role to tests as instruments for the development of theory.

Suppose I claim that I can measure "Strength of the tendency to repress the Oedipus complex." Any psychologist can reject the title of the test and therewith the interpretation, since both are based on what Cronbach and Meehl call a "nomological network" which extends beyond the test validation enterprise. But any evidence for the cohesion of the various items or for the predictive power of the test is public and not contingent on acceptance of psychoanalytic theory. Someone with a different theoretical outlook can reject the proffered interpretation, but he cannot reject at the same time whatever evidence there is of something to be interpreted.

From a practical view, whichever pair of questions is asked, both must be answered. A clinician gets no help from a highly valid test which he cannot interpret. Test theory must make provision for answering both questions. In the remainder of this paper validity will be taken to cover both the extent to which a test measures anything and the validity of its interpretation.

RELATION OF TEST BEHAVIOR TO THEORY

A. TEST RESPONSES AS SIGNS AND AS SAMPLES

Although the view of test behavior espoused here is somewhat different from that of Goodenough [37, Ch. 7], her terminology is useful in describing it. The present view is that responses to items are always and essentially both *signs* and

samples of behav
test responses as
that they repres
presence of trait
which they did r
to test responses
was describing te
essentially simila
is desired to prec
items of every
samples. One wo
unless one could
behavior to beha
ation. This repr
whole reason fo
enough implied
time, test behavi
assumed to obe
determined by
behavior. In thi
have the charact

In principle,
represented in
kinds of actions
cators of certai
as their signals.
sensitive indica
short be called
It seems reason
festations of a
share some of t
festations of th
reference of ev
gotten. The ma
ings of every
about the value
structured, pur

One of the a
is to determine
to find the trait
Unfortunately
cess of this ent
day forms cann
classified. Ther
study of behav
unambiguously
ior can be sol
items. Because
know about b
test behavior.

samples of behavior. In referring to some test responses as signs, Goodenough meant that they represented and indicated the presence of traits and of other behavior which they did not resemble. In referring to test responses as samples, Goodenough was describing tests in which the items are essentially similar to the behavior which it is desired to predict. In a literal sense, the items of every test are both signs and samples. One would not give a test at all unless one could make inferences from test behavior to behavior outside the test situation. This representative quality is the whole reason for giving tests, as Goodenough implied [37, p. 102]. At the same time, test behavior is behavior and must be assumed to obey the same laws and be determined by the same factors as other behavior. In this sense, all test responses have the character of a sample of behavior.

In principle, every trait of a person is represented in each of his actions. Some kinds of actions are such insensitive indicators of certain traits as to be worthless as their signals. If an action is an extremely sensitive indicator of a trait, it may for short be called a manifestation of that trait. It seems reasonable to assume that manifestations of a trait in a test situation will share some of the properties of other manifestations of the same trait. The multiple reference of every action must not be forgotten. The many sources and many meanings of every response induce scepticism about the value of searching for rigorously structured, pure, unidimensional patterns.

One of the aims of psychological theory is to determine the structure of personality, to find the traits which determine behavior. Unfortunately for the simplicity and success of this enterprise, behavior in its everyday forms cannot be uniquely itemized and classified. There are no natural units for the study of behavior. The problem of discrete, unambiguously identifiable units of behavior can be solved by use of objective test items. Because tests are samples, what we know about behavior in general applies to test behavior. Because tests are signs, what we learn from tests can help structure and interpret knowledge of other, more amorphous behavior.

The problem then becomes one of finding items which are sensitive signs of those areas of behavior which are significant for practical and for theoretical purposes. This has proved somewhat easier with respect to abilities than with respect to personality. The present concern is therefore mainly with personality tests, but much of what is said could be applied with slight change to ability tests. What is required might be called a "psychology of objective test behavior." There is needed a theory, supported by data, of what kinds of traits or what levels of personality can be measured by different kinds of personality tests.

B. THE PROBLEM OF HOMOGENEITY

There is a bifurcation in test theory. If the aim of testing is to predict optimally a single criterion, then any item measuring aspects of behavior which are related to that criterion may be included in the test; such items may be independent of each other or even negatively correlated, at least theoretically. Homogeneity of test content is either ignored or deliberately eschewed. If, however, tests are conceived as instruments of psychological theory rather than as devices of psychotechnology, then one has to face the considerable problem of constructing tests which are homogeneous with respect to some trait. The present proposal is that the problem of test homogeneity be viewed in intimate relation with what is known of the complex causation of behavior, and of the fact that traits are manifest in a multiple, alternative, and at times dialectically opposed manner.

A number of attempts to meet the problem of homogeneity have been devised on the basis of a more or less strict logical or mathematical definition of homogeneity. The most important of these is Guttman's [44] scale analysis. A perfect scale is composed of a set of items such that if A has a higher rank than B, then A is as high as or

higher than B on every item. Then for each total score there is just one configuration of item responses. Thus, knowing any individual's score on the test as a whole, one can reproduce each of his answers, a situation obviously different from the usual one in psychological testing. When consideration is restricted to dichotomous items and cumulative tests, that is, tests where each item is scored one or zero and the item scores are added to obtain the total score, a Guttman-type scale is identical with what Loevinger [69] called a perfectly homogeneous test. In this kind of test there is an order of the items such that each individual obtains a score of one on every item up to the one corresponding to his total score and zero on all subsequent items. Scale analysis has come into wide usage in sociology, but the sets of items which prove to be almost perfect scales are usually hardly more than several rephrasings of a single question. They have thus not provided new insights into psychological traits as yet.

Recently Coombs [13] has devised models for conceptualizing completely homogeneous tests. He has propounded a "theory of psychological scaling" which divides psychological tests into two types, which he denotes by Task A and Task B. Tests of ability and expressions of preferences between stimuli illustrate Task A. The judgment of which stimulus has more of some attribute illustrates Task B. In Task A each stimulus and each individual is assumed to have a "scale value"; whether the individual scores plus on an item (agrees, succeeds) depends on whether his scale value is greater than that of the item. The scale value of the individual is assumed to be the same regardless of the item. So far one may be led to suppose that the notion of "scale value" corresponds to such intuitive ideas as ability and trait. However, in Task B the individual is postulated to assume a scale value equal to that of the item as he proceeds from one item to another. Thus it appears that what Coombs calls "scale value" is an arbitrary construct, or set of constructs, which does not correspond closely to intuitive notions about traits.

An interesting feature of Coombs' system, but one which makes it more recondite, is that he emphasizes a formal similarity between the people taking the test and the items, which he refers to as stimuli. Corresponding to Task A is Dual Task A, and corresponding to Task B is Dual Task B. A test falls in the category of Dual Task A if each stimulus (item) classifies the individual as being above or below the item's scale value. A test falls in the category of Dual Task B if each item grades the several individuals with respect to each other on an attribute. In Dual Task B the item itself does not have a scale value. Thus an essay question permits sorting individuals with respect to some attribute, but the question itself does not have a scale value. A problem in arithmetic, however, has a scale value which, in effect, each person taking the test measures himself against (Dual Task A).

Hovland and Sherif [57, 96] have pointed out the inconsistency between Thurstone's method of constructing attitude tests by equal-appearing intervals (Task B) and recent findings concerning influence of personality traits on perception and judgment. They have added data showing the difficulty of isolating the two types of task. Both Edwards [26] and Peak [92] have pointed out that the work of Hovland and Sherif is difficult to integrate with Coombs' separation of Task A and Task B.

Workers in the field of personality measurement have recognized and struggled for years with the fact that self-reports concerning personality traits are subject to such massive systematic distortion as to make them virtually worthless as direct measurements of personality traits. One of the most promising methods for dealing with this problem lies precisely in the direction of straddling what Coombs calls Task A and Task B, stating preferences between stimuli and judging stimuli with respect to an attribute. Campbell [9] has provided an able review of such indirect

attitude measures. His paradigm for the construction of disguised tests of attitudes is:

> . . . a plausible task, (a) which your respondents will all strive to do well, (b) which is sufficiently difficult or ambiguous to allow individual differences in response, and (c) which can be loaded with content relative to the attitude you seek to measure. Test the responses of individuals for persistent selectivity in performance, for correlated or non-random errors [9, pp. 33–34].

The Berkeley F-scale [2] is an example of a disguised test calling for pseudo-judgments (Task B) which in fact reflect prejudices (Task A). Examples of items are: "Most people don't realize how much our lives are controlled by plots hatched in secret places" [2, p. 257]; "The business man and the manufacturer are much more important to society than the artist and the professor" [2, p. 255]. Other items in the F-scale seem to be within the realm of Task A: "There is hardly anything lower than a person who does not feel a great love, gratitude, and respect for his parents" [2, p. 255]. The decision as to whether this item is purely Task A or contains an admixture of Task B is difficult. The distinction is essential to Coombs' theory of scaling, since in one case there is zero correlation between scale values of persons and items, in the other case the correlation is unity. Unfortunately, the scale values correspond to no measurable quantities; so no empirical test is available.

Coombs maintains (personal communication) that such data as those of Hovland and Sherif have no bearing on his distinction between Task A and Task B. His system is apparently so abstract that data have no bearing on his assumptions, and certainly intuition gives no clue as to their validity. One cannot help wondering how such a system can make enough contact with reality to contribute to solution of psychological problems. But Coombs aims, no less than does the present monograph, to integrate psychological and psychometric considerations in his work. Faced with data which do not conform to a unidimensional model, Coombs [15, p. 5] points out that one can proceed to a stochastic (probabilistic) model which allows for error, or one can proceed to account for the data in terms of several dimensions. Coombs' recourse is to multidimensional analysis. But not many more than two or three variables can be handled at a time in this kind of analysis. Coombs has thus equipped us with a kind of analysis which can handle, say, 10 items all of which are completely determined by the same two or three component variables. But where are such items to be found, and how will we know that we have found them? Can we expect anything other than psychological trivialities to turn up in such form?

The problem of making inferences about a single trait from a set of responses all of which are multiply determined is a substantial one. Clinicians, in drawing inferences, are faced with a similar problem. They do not seek aspects of behavior which are determined by a single trait, for there are none. Nor do they seek to analyze all of the many causes of the behaviors they observe. To be confident that one's multidimensional analysis is complete, there must always be many more behaviors observed than there are component causes, and there is no way to insure obtaining any such situation. The clinician searches for a common theme or thread in behaviors which are superficially diverse. When an item of behavior is viewed as an indication of a trait, all other traits and influences which determine it operate as errors. If the observed behaviors are sufficiently diverse, the errors are uncorrelated and more or less cancel each other out.

This solution carries the implication that inferences about individuals will never be made with certainty but will always carry some probability of error. Since that probability of error can, in appropriate circumstances [72], be made arbitrarily small, the objection is not too serious. This solution is what Coombs and others call "actuarial

measurement"; it stays within the main stream of psychometrics, beginning with Spearman.

C. OBSERVATION PRIOR TO MEASUREMENT

Considerations adduced so far have led to rejection of a single external criterion or of rigid or univocal inter-item relationships as the sole touchstone of psychological measurement. Both prediction and structure will recur later in the discussion as components of construct validity, but they cannot stand alone as guides to test construction, prediction because it changes with every slight change in circumstance, rigid structure because such relationships are not found in the most important areas of psychological measurement. Measurement in psychology is a complex process, requiring correspondingly complicated concepts. The theory of measurement being developed herein is germane only to a particular kind of data. It is necessary now to explain that limitation.

The term *objective test* is used in the present monograph in the sense of *structured tests viewed behavioristically.* Objective tests are distinguished on the one hand from projective tests and on the other from rating scales. The former distinction relates to item structure. Projective tests involve free response, as do most interviews, while objective tests, in the sense that the term is used here, require in principle that every individual choose one of the stated alternatives for each item. (Speed tests are usually objective, but they are almost as inconvenient statistically as projective tests and are arbitrarily excluded from the present discussion.) Rating scales usually refer to a different person than the one responding. Self-ratings, when the response is scored categorically, are not excluded here from objective tests. Attitude tests, interest tests, and self-rating questionnaires are alike considered objective tests, as are, of course, power tests of ability, since the responses on such tests are viewed here simply as behavior with no automatic ascription of validity to the apparent content. As Campbell [9] has pointed out, in a review which advocates the use of structured disguised tests, the task of validating disguised measures is identical with that of validating apparently undisguised ones. In view of the well-known deviousness of human nature, the safest course appears to be to assume that all tests are disguised ones. In a direct or undisguised test it is after all only the motives of the investigator, not those of his *Ss*, which are undisguised. The present use of the term *objective tests* has coexisted for many years with a contrary usage which contrasts objective tests and questionnaires.

The requirement of objectivity imposes on us a behavioristic attitude towards the psychological meaning of test responses; equally it implies restrictions on the psychometric meaning of responses. Although many of the considerations of the present paper may apply more widely, the discussion will be directly concerned only with dichotomous items for reasons which are by no means arbitrary or trivial. Most of test theory, indeed, has been worked out for dichotomous items, but the justification, if any, has been chiefly statistical simplicity and convenience.

An alternative to dichotomous items is Likert-type items. For such items an extreme statement is presented together with a rating scale, often having between four and seven points. *S* may be required to check, for example, one of the following, "Strongly agree," "Agree," "Disagree," or "Strongly disagree." Very often arbitrary scores will be assigned to the alternatives, from 4 for "Strongly agree" to 1 for "Strongly disagree." The scores on the relevant items are then cumulated to form the total score. This procedure implicitly assumes that the difference between agreeing strongly and simply agreeing is the same as the difference between agreeing and disagreeing, an assumption which is neither reasonable nor in accord with what is known about the independence of intensity and direction of conviction. It has been

found repeatedly that some individuals have a tendency to express or accept extreme opinions as such, regardless of the topic. This fact introduces a spurious correlation between items and between tests composed of Likert-type items.

Likert-type items have at times been defended on the grounds that the intensity component is valid variance for some purposes. But, in principle, all components of test variance must be valid for some purposes. The problem is that in life the various components of variance are confounded. The task of psychometrics is to isolate, to identify and, so far as possible, to measure separately the important components of variance. Likert-type items, or any other test techniques which deliberately confound various sources of variance, operate in the wrong direction.

McQuitty [83, 84] has at times handled three-choice items by treating each of the three responses to an item as if it were a separate item. In this case the assumption that the three alternatives lie on a single continuum is avoided, but other difficulties are introduced. For two such three-choice items there will be nine coefficients of correlation, some of them probably based on very small numbers of cases. Moreover, the nine coefficients will not be experimentally independent; choosing one of the three alternatives excludes the possibility of choosing one of the other two. Thus no two of the nine coefficients can be introduced into the same correlation matrix for analysis by usual methods.

In many tests responses are recorded in several categories but are scored dichotomously. This procedure appears to be the general rule in scale analysis [100, 107] where it leads to special difficulties which will be discussed under Components of Construct Validity, *B*, below. Where several different responses are given a single score, the subsequent analysis does not deal with responses as raw data but rather with what MacCorquodale and Meehl [81] have called an intervening variable. Indeed, what are called "raw scores" in traditional psychometrics are summaries of several responses rather than the original responses and therefore are intervening variables. (The exceptional case where to each score there corresponds just one configuration of responses is what Guttman calls a perfect scale.) Since intervening variables are not responses, they do not in general have the characteristics of individual manifestations of a trait. For example, two scores which relate to a single trait would ordinarily correlate more highly than two responses which manifest the same trait because scores are more reliable than single responses. The distinction between responses and scores is crucial to the present view. The basis of the psychometrics advocated here is analysis of responses prior to analysis of scores. Thus psychometrics remains imbedded in psychology.

Restriction of the present consideration to dichotomous items is not a commitment to the check-list form of personality test, where the individual simply checks those items he likes or agrees with. Coombs [14] refers to such items as "irrelative"; items which involve a choice among alternatives he refers to as "relative." Dichotomous items can also be relative; that is, an individual can be presented with a forced choice between a pair of alternatives. In this instance one of the alternatives, say, the first, can be thought of as chosen or not chosen. No information is lost and no spurious information is introduced in the dichotomous representation of such items.

When Likert-type items are scored as continua, each item is in effect considered a little measurement. With dichotomous items, however, it is possible to think of each item not as representing a measurement but merely as an observation. A certain bit of behavior, the positive alternative, is either present or absent. The individual either checked the first alternative response or he did not. Most expositions of psychometrics, it is true, discuss dichotomous items as imperfect representations of underlying traits which are conceived of as continuous and often as normally distributed.

The present paper will endeavor to show how a theory of measurement can be based on dichotomous items conceived as observations without introduction of any surplus meaning of measurement into the individual items.

In an early monograph on attitude measurement Thurstone [106] distinguished two types of tests, *increasing probability* and *maximum probability* tests. Thurstone's method of test construction by equal-appearing intervals led to a maximum probability type of tests; each item is assigned a scale value and an individual's score is some average of the scale values of the items on which he scores plus. Ordinary tests of ability are increasing probability tests. The two types of tests were often confused, as pointed out by Loevinger [70], who proposed the terms *cumulative* and *differential* for the types of test. Coombs [14] has generalized the distinction somewhat with the terms *monotone* and *non-monotone* items. A monotone item is one for which an increase in the amount of the underlying trait will never decrease the probability of answering plus on the item. Non-monotone items are not excluded from the present discussion. But Thurstone-type tests, which assign a scale value to items as a means of computing a score, have a surplus meaning of measurement attached to items, just as Likert-type items do. They are therefore excluded from the discussion. Some of the ways in which combinations of dichotomous items can generate actuarial measurement of traits are discussed under Components of Construct Validity, B, in connection with the structural component of validity.

The present monograph is, in summary, restricted to consideration of dichotomous items viewed as present-absent observations of behavior whose psychological referents must be established by evidence. The alternatives lead either to injecting psychological or psychometric surplus meaning into the observations or to combining several observations into a single intervening variable. In the former case objectivity is surrendered, whereas intervening variables lose some of the character of samples of behavior.

D. THE PSYCHOLOGY OF OBJECTIVE TEST BEHAVIOR

Let us return to the psychology of objective test behavior. The naive assumption that the trait measured by a personality test item is always related in a simple, direct way to the apparent content of the item has long since been disproven. Content, of course, is not a negligible factor, but content mediates the traits brought into play by an item in a more subtle and indirect fashion than early workers (and some present ones?) believed. Interaction of type of item with area of content in the setting of the objective test situation must be reckoned with.

A model of this kind of theorizing is Schafer's essay [95] on psychoanalytic interpretation of Rorschach responses. One cannot be willing permanently to settle for pure theory without rigorous empirical checks, even though agreeing with Schafer that there are methodological difficulties in the verification of his hypotheses. Nevertheless, his book is a valuable example of how theory, in his case, psychoanalytic theory, can generate hypotheses about the relationship of various traits (drives, defenses, and adaptive qualities) to various aspects of test behavior.

Although less searching from a psychodynamic view than Schafer's book, an earlier paper by Lindzey [68] concerning the Thematic Apperception Test illustrates the integration of theory with data as well as the integration of psychometric with psychological considerations. Such a review of objective test behavior is needed. Undoubtedly many papers focused on other topics can be made to yield data for such a review.

Meehl's paper [85] on the dynamics of structured personality tests represents a noteworthy milestone. He contrasted two approaches to such tests. In the first S is asked about his behavior as a substitute for observing the behavior. Frequently as-

sociated with this approach is *a priori* construction of scoring keys, "requiring the assumption that the psychol ist building the test has sufficient insight into the dynamics of verbal behavior and its relation to the inner core of personality that he is able to predict beforehand what certain sorts of people will say about themselves when asked certain sorts of questions" [85, p. 297]. The second approach "consists simply in the explic denial that we accept a self-rating as surrogate for a behavior sample, utes the assertion that a 'self-ra tes an intrinsically interesting t bit of verbal behavior, th lates of which must be dis rical means" [85, p. 297].

At the time Meeh was probably true that *a p* disguised scales were a psychologists did not k bout objective test behavior vance the nature of disguised ssions. Accumulated experience ast ten years would lead one to s partially the question of whether items are disguised and whether their significance can be guessed in advance by an astute psychologist. Meehl's example of nonsomatic items in the *Hy* scale of the Minnesota Multiphasic Personality Inventory (MMPI) illustrates just this point. These items, he points out, all seem to say, "I am psychiatrically and socially well-adjusted." Since these items as a group are good indications of hysteria and hysteroid temperament, they are not valid if taken as direct representations of behavior. When viewed as indications of lack of insight, repression and dissociation, and the *belle indifference* of these patients, the items are seen to possess a theoretical relation to the syndrome which, empirically, they signify. Presumably today the meaning of such responses is more likely to be disguised from patients than from test constructors; but this statement is not to be taken as advocacy of *a priori* construction of scoring keys.

Meehl cites as examples of the empirical approach the MMPI and the Strong Interest Blank. While the Strong test was, indeed, constructed empirically, it has not been tied primarily to traits but to external predictions. A more important predecessor of the MMPI is the M-F test of Terman and Miles [101], which has been almost lost in limbo. Far more of contemporary personality measurement derives from their ingenious set of studies than is currently acknowledged. In particular, the principle that the possibility of measuring a trait accurately may depend on S not knowing what the investigator is measuring was dramatized, if not introduced, in their work. E. L. Kelly [62] participated in experimental demonstration of the principle. Campbell [9] cites some earlier but less impressive uses of disguised tests.

A few further references may clarify the purview of the psychology of objective test behavior. Meehl [85] observed that while ambiguity in wording and inaccuracy of memory are sources of error in the traditional view of self-ratings, for the MMPI they may be sources of discrimination. Elias [27] found evidence that ambiguously worded items served projective purposes better than explicitly worded ones. Dorris, Levinson, and Hanfmann [22] found some evidence that third person items are better measures of defended against or unrecognized personality tendencies than first person items. Owens [91] found that validity was raised markedly by changing the format of a test from that of choosing or not choosing each of 30 statements to that of forced choice between 27 pairs of more or less contrary statements. The second (paired choice) test included a number of items taken directly from the first test, and these were, in fact, the most valid items of the second test. Some, but not all, more recent studies have confirmed the advantage of forced choice items [34].

Let us suppose that we have made some progress in finding combinations of types of items and areas of content which are reasonably sensitive indicators of basic personality traits. Clearly, theory is needed to find such items. But how can theory in turn be enriched by them? Objective test

items are uniquely accessible to study in their relations to each other, in the distribution of responses in the population, and in their relations to outside variables. Such data have much to offer in clarification and extension of psychological theory.

COMPONENTS OF CONSTRUCT VALIDITY

The APA Committee on Psychological Tests divided validity into four types, content validity, concurrent validity, predictive validity, and construct validity, with the implication that it is optional which kind of validity is proposed for a test [121]. With respect to construct validity, all of the other kinds of validity are possible supporting evidence, but again there is an implication of option. This analysis of validity was accepted by Cronbach and Meehl. The categories were not claimed to be logically coordinate; the fact that they are not is nonetheless disturbing.

The construct validity of a psychological test, that is, its validity as a measure of real traits, is in the present monograph conceived as having three aspects: the substantive component, structural component, and external component. These three aspects are mutually exclusive, exhaustive of the possible lines of evidence for construct validity, and mandatory. The substantive component is somewhat different from what was previously called content validity. The structural component includes, but is not exhausted by, such concepts as homogeneity and functional unity. The external component includes relation to non-test behavior, factorial pattern or relation to other tests, and absence of distortions. The predictive and concurrent validities of the previous papers are an alternative subdivision of the external component.

Thurstone [103, Ch. 14] proposed similar categories of evidence for the validity of factors. The suggestion that tests have internal and external validity has come from many sources, including Cattell [11, p. 545], Peak [92], and Guttman [47, p. 57]. Internal validity is here divided into substantive and structural components, again following many others.

The three aspects of validity are closely related to three stages in the test construction process: constitution of the pool of items, analysis of the internal structure of the pool of items and consequent selection of items to form a scoring key, and correlation of test scores with criteria and other variables. Any test composed of items contains an explicit or implied commitment with respect to each of these steps. Some items were considered for inclusion and some were not. Of those considered, some were included and some were not, often but not always on the basis of structural considerations; the nature of the scoring key, in any case, contains an implied commitment to a particular structure [92]. And a test cannot be presented seriously for clinical use unless something is known of its external correlations, possibly leading to modification in the test or in its interpretation. None of these steps in test construction is optional, and none is without consequence for the validity of the test in use.

A. SUBSTANTIVE COMPONENT

1. Use of content in item selection

There are many methods which have been used to select items to form a test. They can be classified according to whether the chief criterion is internal consistency, external validity, or reasonableness of content. Perhaps all major test construction projects have used a combination of criteria, but generally they have relied predominantly on a single one. The discussion in this section does not seek to evaluate these methods as a whole but simply to show how content is utilized in the several methods.

a. Content validity

According to the *Technical Recommendations*, "Content validity is evaluated by showing how well the content of the test samples the class of situations or subject

matter about which conclusions are to be drawn" [121, p. 13]. Concern with content validity is appropriate when one wants to know how S will perform "in a given universe of situations of which the test situation constitutes a sample" [121, p. 13].

Clearly, tests of educational achievement have most often been constructed with a concept similar to content validity as their justification. Achievement tests, as such, are not within the scope of the present monograph; nothing said herein is intended as criticism of methods of constructing such tests. Guttman, however, has for some years advocated content validity in the context of tests as measures of traits.

Using the term *attribute* to mean item (implying, item as observation rather than as miniature measurement), Guttman states, "An attribute belongs to the universe by virtue of its content. The investigator indicates the content of interest by the title he chooses for the universe, and all attributes with that content belong in the universe" [44, p. 141].

Guttman's emphasis on the investigator's decision as to the nature of the universe and whether any particular items fall in it is the justification for the charge made previously under Extension of the Concept of Validity that the argument of content validity is *ad hoc*. To change the investigator would ordinarily be to change at least somewhat the definition of the universe and the delimitation of the items within it, as Goodenough [37, p. 104] pointed out. Studies by Ruth Tolman, Grayson, and Forer [31, 38] illustrate this point. Clearly, however, the *Technical Recommendations* and some other expositions of content validity seek to avoid this arbitrariness. The *Technical Recommendations* list as essential criteria for content validity:

> If a test performance is to be interpreted as a sample of performance in some universe of situations, the manual should indicate clearly what universe is represented and how adequate the sampling is. . . . The universe of content should be defined in terms of the sources from which items were drawn, or the content criteria used to include and exclude items. . . . The method of sampling items within the universe should be described [121, p. 20].

The problem which arises is that the more one objectifies the nature of the universe from which the sample of items is to be drawn, the less likely is the universe to represent exactly the trait which the investigator wishes to measure. Moreover, for any given trait name, two investigators would not necessarily specify the same objective domain from which to draw a sample, nor the same method of sampling. Should a vocabulary test be drawn from an unabridged dictionary, an abridged dictionary, or a list of words in common use? And which one? One must either decide arbitrarily in favor of one such alternative and run the corresponding risk that another investigator might make a different decision, or one must demonstrate that the choice is inconsequential. To accomplish the latter requires a great deal of data and thus introjects strong considerations other than those of pure content.

A work sample test would seem to provide an optimal instance of content validity. Consider again, vocabulary. Certainly every word which can be defined correctly is a part of vocabulary. But Tucker [109] found that certain words had very low relation to the test as a whole, despite the fact that most words had a high relationship to vocabulary as a whole. The "poor" words had some common characteristics, but they could not have been identified in advance; they were not technical words. Apparently one will do a better job of measuring vocabulary as a whole, vocabulary as a trait, by omitting than by including the aberrant words. But to do so is to admit considerations other than content.

Both the authors of the *Technical Recommendations* and Guttman consider essential the evaluation of the internal consistency of the set of items considered to measure a given universe of content. Guttman envisages, and the *Technical Rec-*

ommendations do not exclude, possible elimination of a small percentage of items if they do not hold together with the rest of the test. Thus some utilization of empirical evidence is regularly expected even when the chief basis of test construction is content.

"The heart of the notion of content validity" [67, p. 299] is that the test items are samples of the trait-universe to which generalization will be made. In the great majority of psychological tests the behavior sampled is verbal and the trait to which generalization is made includes much nonverbal behavior. Thus content validity is not applicable to such tests. But the point can be pushed further. Test behavior is always utilized as a sign of non-test behavior, as emphasized previously under Relation of Test Behavior to Theory. The very fact that one set of behaviors occurs in a test situation and the other outside the test situation introduces an instrument error which is ignored by the concept of content validity. The psychodynamics of testing, "the psychology of objective test behavior," can never safely be omitted in drawing inferences from test behavior. The graduate student who performs most brilliantly on his qualifying examinations will not necessarily be the most brilliant professor, nor will the one who becomes amnesic on his examination necessarily be equally blank under less stressful circumstances. More strikingly, the person who is best able to respond in a test situation in a warm and friendly manner is not necessarily the warmest and most friendly person; it is quite conceivable that there are traits such that the capacity to simulate them indicates their absence rather than their presence.

There is a continuum of tests ranging from those whose content is most similar to the behavior it is desired to predict to those whose content least resembles the behavior to be predicted. The problem introduced here is the problem of the disguise of measurement. Considerations of content are most fruitful for theory, however, precisely in those cases where the test is most disguised, where the content bears the least apparent relation to the trait. The term *substantive validity*, which will be defined presently, is introduced in the present monograph because of the conviction that considerations of content alone are not sufficient to establish validity even when the test content resembles the trait, and considerations of content cannot be excluded when the test content least resembles the trait.

b. Empirical keying

At the opposite extreme from tests justified solely in terms of reasonableness of content are those justified only in terms of empirical properties of items. The term empirical keying usually refers to selection of items according to their correlation with an external criterion, a method whose deficiencies have been explored previously under Relation of Test Behavior to Theory. It might as well refer also to selection of items according to a criterion of internal consistency, such as factor analysis. At this point in the discussion the distinction between external and internal empirical criteria will not be important. Examples of empirically keyed tests are the MMPI, which began with a collection of all personality test items of a certain type then in use, and biographical inventories, where even more heterogeneous items are included in the original test form. A particular key for such a test will often involve only a small proportion of the items.

In the case of many empirical keys which are used over a period of years, for example, those of the *Strong Vocational Interest Blank*, little or no attempt is ever made to examine the content for common themes in a given key. One suspects that some ultra-empiricists would consider such examination of content as immoral or unscientific. Yet an experimentalist is not considered more scientific if he collects data and walks away from it without seeking explanation of the behavioral dynamics which account for it. If theory is fully to profit from test construction as a part of psychology, every

item included in a scoring key must be accounted for; a less strong case can be made for explaining the exclusion of items [cf. Jessor and Hammond, 60].

Even among the test constructors who are interested in examining the content of empirical keys, it would almost certainly be considered gross scientific impropriety to delete items because no reasonable connection could be found between them and the content chiefly indicated for the test. This attitude is the opposite of that indicated by the term content validity; yet the criteria for choosing one view rather than the other are not easily applied to a particular situation.

The dangers of pure empiricism must not be underestimated. There is always a probability that an item be included in a scoring key by chance; the probability can be reduced, at considerable cost, but it cannot be made zero. Quite apart from inclusion of items by chance, items may be included because of fortuitous but misleading correlation with the trait measured. Travers [108] has worked out a striking example of the dangers of pure empiricism in the absence of hypotheses.

There appears to be no convincing reason for ignoring content nor for considering content alone in determining the validity of a test or of individual items. The problem is to find a coherent set of operations permitting utilization of content together with empirical considerations.

2. *The universe and the pool*

The first step in test construction is constructing or (often) collecting a pool of items. Some decision, broad or narrow, is made as to content at that point. "I wish I were not so shy" was not considered for inclusion in the pool of items from which the Stanford-Binet was constructed; problems in long division were clearly inappropriate for inclusion in the MMPI. A test constructor must have some purpose in mind, and this purpose defines the universe of discourse or area of content from which he will choose items for the original pool. The primary purpose may be to improve an external prediction, to measure a trait postulated by theory, or to investigate the structure of some aspects of behavior. (Whichever is the investigator's chief purpose, he is ultimately involved in all three enterprises if he wishes to relate his test to theory.)

An unfortunate confusion between the pool of items and the area of content has arisen in test theory [74]. This confusion is signalized by the term "universe of items." Papers on test theory often assume that items are chosen randomly [77] or simply independently and without regard to empirical properties [48] from a universe of items possessing a characteristic such as some form of homogeneity. In practice, tests are usually constructed by choosing the best items from the pool used in the test construction research. While the pool is chosen on an *a priori* basis, the choice of items from the pool is made according to the properties of the items revealed by empirical study.

The point to note is the difference between what actually takes place in constructing a test and the process implicitly assumed in test theories. Random selection from a "universe of items" does not describe selection of the pool of items or of the scoring key. So far as constitution of the pool is concerned, selection of a single item automatically excludes many others which differ only slightly from the chosen one. Thus several items are neither randomly nor independently selected. The term "universe of items," moreover, obscures the fact that between the presumably unlimited number of items representing a given content and the finite pool of items actually studied there intervenes an idiosyncratic, nonreproducible process, the process by which the given investigator or group of investigators constructs or selects items to represent that content. Although this process, the constitution of the pool of items, is a universal step in test construction, it has not been adequately recognized by test theory. But this step in test construction is crucial for

the question of whether evidence for the validity of the test will also be evidence for the validity of a construct.

Cronbach and Meehl point out that in any characterization of a cluster of items

> . . . 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 [20, p. 292].

The reasoning of the above quotation is not, I believe, open to question. Guttman has made similar observations about the successive redefinition of "universes" according to whether predictions are verified. Logically, of course, predictions are indistinguishable from concurrent correlations. It follows directly from the above quotation that to characterize a cluster of items with any confidence requires a knowledge not only of a number of items included in the cluster but also of many items which fail to meet requirements for inclusion in the cluster. Indeed, the excluded items might reasonably be required to exceed greatly the items included. The present proposal is that this requirement be met not in a lengthy series of investigations, though verifying studies are always in order, but primarily in the very constitution of the pool of items from which the test is chosen.

The use of tests for the elucidation of constructs suggests the following principles, in descending order of generality:

At very least, the items in the pool should be drawn from an area of content defined more broadly than the trait expected to be measured.

When possible, the items of the pool should be chosen so as to sample all possible contents which might comprise the putative trait according to all known alternative theories of the trait. Thus the empirical data utilized for construction of the scoring key simultaneously test hypotheses about the underlying trait. This principle has been advocated and practiced by Thurstone [104] and Guilford [40, 42].

The proportion of items of different content is not specified by the previous principle, but may affect the outcome of factorial studies. A principle which suggests itself here is one (or perhaps a modification of one) proposed by Brunswik [7]: that *the various areas or sub-areas of content should be represented in proportion to their life-importance.* This suggestion is similar to one made by T. L. Kelley [61] and by Cattell [11, p. 215], who assumed that life-importance could be judged by dictionary representation.

A recent proposal by Guttman[1] seems to offer an alternative to representation of areas of content according to life-importance. Stephenson [99] earlier advocated a similar design for Q-sort items. Guttman proposes that what might be called the logical dimensionality of an area be predetermined by the investigator; e.g., a set of items may differ in form, content, complexity, and other dimensions. These dimensions correspond to Fisher's "factors" in analysis of variance designs. Guttman proposes to call them "facets" to avoid confusion with factor analysis. Each facet has a set of values; for example, the content possibilities might be words, numbers, and geometric figures, the forms multiple choice, true-false, and completion, and so on. The experimental design Guttman proposes is that each value of each facet be paired with each value of every other facet,

[1] Dr. Guttman has presented a number of papers on facet analysis which have not as yet been published. This discussion is based on a mimeographed paper by him, "What lies ahead for factor analysis?" issued from the Center for Advanced Study in the Behavioral Sciences, Stanford University, 1956.

items of the pool
s to sample all pos-
night comprise the
; to all known alter-
ait. Thus the empiri-
construction of the
usly test hypotheses
trait. This principle
l practiced by Thur-
rd [40, 42].
ms of different con-
the previous prin-
t the outcome of
ciple which suggests
perhaps a modifica-
by Brunswik [7]:
or sub-areas of con-
ented in proportion
. This suggestion is
y T. L. Kelley [61]
215], who assumed
ould be judged by
on.
Guttman[1] seems to
o representation of
ding to life-impor-
earlier advocated a
ort items. Guttman
ight be called the
of an area be pre-
estigator; e.g., a set
form, content, com-
nensions. These di-
Fisher's "factors" in
signs. Guttman pro-
cets" to avoid con-
sis. Each facet has a
ple, the content pos-
ords, numbers, and
ne forms multiple
completion, and so
lesign Guttman pro-
ne of each facet be
of every other facet,

ited a number of papers
ve not as yet been pub-
sed on a mimeographed
ahead for factor anal-
ter for Advanced Study
s, Stanford University,

as in an analysis of variance design. The intention apparently is to combine this experimental design with a new form of factor analysis which Guttman calls radex theory [50]. Until further details and examples are published, it is not possible to evaluate the proposal as a whole. The following comments are meant only to apply to facet analysis in present context, not in all the many applications which Guttman foresees for it.

Logical analysis of an area of content on some such basis as facet analysis, if supplementary to the principles proposed above, is valuable to insure complete coverage of the area and has been practiced by psychologists professionally engaged in test construction and analysis for many years. As an alternative principle to representation in terms of theories at stake or in terms of life-importance of sub-areas, it is not acceptable. Pairing every value of every facet with every value of other facets may at times be a highly artificial procedure. Some contents may adapt themselves far more readily to some forms than others. Moreover, important sub-areas of content may be missed by this design. For example, problems may involve words or numbers or words and numbers. It may make sense for several values of a single facet to be represented in a single item. This possibility violates the neat but artificial Fisherian design which Guttman proposes. Brunswik's [8] criticism of factorial design is particularly apposite to Guttman's facet analysis and Stephenson's similar proposal.

It should be noted that neither Stephenson nor Guttman uses facet analysis exclusively. The examples in the foregoing exposition are those of the writer and not of Stephenson or Guttman. The point is that these are simply elaborate and complicated examples of content validation, of *a priori* definition of constructs. Neither Guttman nor Stephenson advocates the objective delineation of constructs through observation of the variety of items included in and excluded from the cluster defining the construct in the manner envisaged in the present monograph.

3. *The concept of substantive validity*

The proposal of this monograph is that the pool of items be constituted on the basis of some broad area of content. The empirical relations among the items, and perhaps between items and criteria, serve as basis for selection of items from the pool to form a scoring key. If the pool has been properly constituted, the process of test construction also tests a family of hypotheses about the trait to be measured. The principles by which the pool of items is constituted have been discussed above; the next section will discuss some principles for the selection of items from the pool. The *substantive component* of validity is the extent to which the content of the items included in (and excluded from?) the test can be accounted for in terms of the trait believed to be measured and the context of measurement. Context includes psychological theory and, in particular, "the psychology of objective test behavior."

The program implied by content validation is inclusion of items in a test solely on the basis of theory; empirical keying implies inclusion of items in a test solely on the basis of data. The present program is that items be included in the original pool on the basis of a judgment of relevance to a broadly defined field. The final selection of items shall be made, however, on the basis of empirical findings. The substantive component of validity is the ability of theory to account for the resultant content; it cannot be determined unless the pool of items is broader in scope than the test.

B. STRUCTURAL COMPONENT

1. *The concept of structural validity*

The *structural component* of validity refers to the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured. This concept will seem novel to many psychologists, but undoubtedly others will find it congenial or even familiar. The concept of structural

validity includes both the fidelity of the structural model to the structural characteristics of non-test manifestations of the trait and the degree of inter-item structure. Given fidelity of item structure, the more highly structured test may be said to have greater structural validity. Since previous discussions of structure have put more emphasis on degree of structure than fidelity of structure, the present paper will reverse the emphasis. Previous reviews of structural theory include chapters by Peak [92] and Coombs [14].

The concept of structural fidelity is based on the ideas developed previously under Relation of Test Behavior to Theory. By confining consideration to dichotomous items considered as observations rather than as miniature measurements, we have insured that the test responses can be considered legitimate samples of behavior. Since test behavior is a sample of behavior in general, it may be assumed to share the characteristics of other behaviors, in particular, the structural characteristics. If the analysis were based on scores rather than on the original item responses, those structural characteristics might well be obscured.

For any given trait it seems reasonable to assume that there is an upper limit to the intercorrelation of its manifestations, which might be called its characteristic intercorrelation. For example, two manifestations of numerical ability would be expected to be more closely related than, say, two manifestations of aggressiveness. Two manifestations of verbal facility would be more closely related than two manifestations of introversion. It is necessary to add, however, that the characteristic correlation is an upper limit and that the lower limit is always zero, as one proceeds from those actions most determined by the given trait to those least determined by it.

According to this line of thinking, the same characteristic value would define the upper limit of correlation of two non-test manifestations of a trait, of two items measuring that trait (ignoring distortions produced by the test situation, which would raise the inter-item r), and of an item with a non-test manifestation. If this view proves correct, such inferences as can be drawn from test behavior will not be based on direct correspondence between individual items and particular behaviors outside the test situation, except in so far as both are related to a central trait. (Cf. Lazarsfeld's definition of a pure test: "All interrelationships between the items should be accounted for by the way in which each item alone is related to the latent continuum" [65, p. 367].) The injunction against interpretation of individual items is not, of course, a new one, but the present reason is somewhat different from the usual one. It is unfortunate, however, that the weakness of item responses as measures of extra-test behaviors led most psychometricians to concentrate on total scores almost to the exclusion of individual responses. The problem of the structural relations of responses was thereby obscured.

Sociologists have in recent years shown great preference for the structural model of Guttman-type scales. Studies by Riley, *et al.* [93], and some other users of Guttman scales appear to be predicated on the assumption that the most rigorously structured mathematical model is invariably the preferred one. This view is contrary to the view proposed here, that the outstanding virtue of a model is its fidelity to what is known about manifestations of the trait or type of trait involved.

The question arises, how can items be found which have at least apparently a closer or more rigorous relation to each other than is in fact characteristic of manifestations of the given trait? There is no unique coefficient of correlation for two dichotomous items. Some coefficients are prejudiced in favor of and some against items which differ in difficulty. In order to define uniquely the characteristic correlation of the manifestations of a given trait, further specifications must be made, such as "phi coefficient for two items each of which characterizes half the population." The Guttman-scalability of a set of items

can be raised
which are widel
analog, rather t
near median diff
scales are usu
same thing is ac
cutting points v
are reduced to
technique of sca
[71] has shown
spaced in diffi
lowers validity
istic item correl
most often the c
statistical rathe
extraordinarily l
raising validity
considering the
analysis.

Guttman, ind
provement of
items, at least a
argued[2] that su
puts the inves
Spearman, who
in favor of a s
basis of a set of
those found to l
Thus Guttman
to determine th
acteristic of wl
rather than pr
precise measur
precise trait. In
is closer to that
which advocate
characteristics
his own disci
mizing multiple
mize scalability
a practice wh
since this pract
of items.

A Guttman-t
sponses of the
knowledge of t

[2] Mimeographe
the development
Center for Adva
Sciences, Stanfor

can be raised by choosing those items which are widely spaced in difficulty or its analog, rather than those items clustered near median difficulty. That, in fact, is how scales are usually produced [100]. The same thing is accomplished by selection of cutting points when multiple choice items are reduced to dichotomies in the Cornell technique of scale analysis [46]. Loevinger [71] has shown that while selecting items spaced in difficulty raises scalability, it lowers validity for tests whose characteristic item correlation is not high, which is most often the case. While the argument is statistical rather than empirical, there is extraordinarily little empirical evidence for raising validity by improving scalability, considering the amount of interest in scale analysis.

Guttman, indeed, does not sanction improvement of scalability by selection of items, at least as a general practice. He has argued[2] that such use of the scale model puts the investigator in the position of Spearman, who was criticized for arguing in favor of a single general factor on the basis of a set of tests "purified" by deleting those found to have "overlapping specifics." Thus Guttman favors use of scale analysis to determine the degree of scalability characteristic of what he calls "the universe," rather than production of an artificially precise measure of an intrinsically nonprecise trait. In this respect Guttman's view is closer to that of the present monograph, which advocates fidelity to the structural characteristics of the trait, than to that of his own disciples. However, in dichotomizing multiple choice items so as to maximize scalability [46], Guttman is permitting a practice which he otherwise eschews, since this practice is equivalent to selection of items.

A Guttman-type scale exists when all responses of the *Ss* can be reproduced from knowledge of their total scores; this is the definition of a scale. Then in substituting scores for patterns of item responses, no information is lost. So bemused have sociologists been by the demonstrated virtues of true Guttman-type scales that other devices in addition to selection and dichotomization of items have been used for artificial production of scales. Stouffer, Borgatta, Hays, and Henry [100] have used "contrived items," each composed of several items; scalability is then sought for the contrived items rather than the original responses. Guttman [49] has proceeded in a different direction, assigning to each nonscale response pattern a corresponding scale pattern, by a method called "image analysis" or the "Israel Alpha technique for scale analysis." Neither method retains all the information in the original data, which was the virtue of a true scale, and neither method has been shown to maximize any desirable quality of a test. Stouffer's method is shown to be slightly better than older techniques of item selection based on scale analysis.

Occasionally recognized but much more often forgotten in the circles where scale analysis is practiced is the distinction between the discovery of structure in a set of items and the imposition of structure upon the items. If tests are to serve as instruments of theory, no consideration is more crucial. Guttman [44] early took a stand against empirical item selection altogether, since he believed the function of structural analysis to be discovery rather than creation of structure. Nonetheless, he has sanctioned many methods which, in effect, create structure.

Coombs has been far more perceptive with respect to this point [13, 14], repeatedly and in various contexts stressing the difference between a structure which may or may not appear in a set of items and a structure which the methods employed cannot fail to reveal. The model of a perfect scale was one which a set of items could or could not conform to; presumably in practice the conformity was never perfect. On the other hand, some kinds of

[2] Mimeographed paper, "A personal history of the development of scale analysis," issued from the Center for Advanced Study in the Behavioral Sciences, Stanford University, 1955.

image analysis cannot fail to produce a perfect scale. Thurstone's method of constructing attitude scales by equal-appearing intervals can hardly fail to produce a scale with equal-appearing intervals. Yet Thurstone's method need not and Guttman's analysis does not involve selection or rejection of items. The writer, in connection with published research [23], has systematically searched pools of as many as 50 items and failed to find any combination of items for which Kuder-Richardson Formula 20 revealed appreciable homogeneity, say, .4. Thus, the question of whether structure is imposed or discovered is different from whether item selection is permitted.

The present monograph proposes that items be selected from a large pool on the basis of empirical properties, in particular, that those items be selected which best conform to an appropriate structural model. The method should be such that items which conform to the model may or may not appear. Ordinarily the degree of structure of the resulting test should also be assessed.

Selection of items does raise a somewhat different problem, that of cross-validation. Tukey states: "In every field of science, and particularly in fields where data and analysis [are] complex, there are two different phases of quantitative analysis—exploration and confirmation—and almost always, when dealing with *complex* problems, these have to be carried out on *different* samples of data" [111, p. 66]. Thus if a given set of items is discovered to have structural coherence in one set of data, the hypothesis of structural coherence must be tested with a fresh set of data.

2. *Some kinds of structure*

What kinds of structure are known to exist among the manifestations of various traits? An exhaustive list of possible structures will not be attempted here; such lists quickly sink to a lifeless and pedantic level.

a. Quantitative models

The most obvious kind of structure is the one in which the number of manifestations is an index of the amount of the trait. Such a trait is appropriately measured by a cumulative test, that is, a test whose score is the number of items marked plus. This structure is the one which has been most adequately explored statistically. Different kinds of structure have the cumulative or additive character, the difference lying in the degree of relationship among the items.

When the degree of relationship among items is low or moderate, optimal construct validity is produced by selecting "median equivalent items," i.e., those items having highest intercorrelations and lying closest to 50% difficulty or its analog ("marginal frequency"). This is the model which has been most thoroughly exploited in psychological test theory and will here be called the classical quantitative model. The method of Loevinger, Gleser, and DuBois [75] for constructing homogeneous subtests from a large pool of items has been evolved explicitly for this type of data. Items are selected for their ability to maximize a coefficient equivalent to the Kuder-Richardson Formula 20, which is considered currently the best measure of test homogeneity [121].

When the intercorrelation of the items is very high, spacing the items in difficulty improves validity. What constitutes "high" depends on the number of items and other factors [71]. Guttman's scale model [44], which is similar to a model for a perfectly homogeneous test proposed by Loevinger [69], is appropriate for this case. Unfortunately Guttman does not publish the relationship between the computational techniques he recommends and the brilliant mathematical derivations from which they derive their prestige. This observation is conspicuously true of his coefficient of reproducibility, which has often been criticized. His recently proposed methods of image analysis are ostensibly based on image theory [48], but the intermediate steps in the reasoning are not available in accessible publications.

Guttman's scale model and Loevinger's

model of a homogeneous test are discussed more extensively by Peak [92], who criticizes them on grounds which appear to be similar to what is here called structural fidelity. Humphreys [58] and Carroll [10] have made similar criticisms.

Tucker [109] has recently worked out a model for cumulative tests intermediate between the classical quantitative type and the Guttman-scalable type. Most words in a long vocabulary list conformed to his model. The model may be described as follows: the items are assumed to be homogeneous in the sense that all correlation between them is accounted for by a single ability, as in Lazarsfeld's latent structure analysis. The scale of ability is so chosen that the regression of proportion passing the item on ability is described by approximately the same ogive for all items, except that the ogive varies parallel to the scale of ability. The shifting of the ogive for different items corresponds to changes in difficulty level. Items are dispersed in difficulty, like the Guttman scale and unlike the classical quantitative model, which concentrates difficulties near the 50% mark. The relation between an individual's total score and his score on any item is probabilistic rather than almost certain, as in Guttman scales. Perhaps Tucker's model represents the closest approach to Guttman scales which can be expected with psychological test materials.

Methodologically, working out a structural model in conjunction with a test construction problem which is intrinsically interesting appears highly desirable. Empty models, models which apply to no content, are thereby guarded against. Tucker's type of scalability may therefore be expected to find further use. For most purposes, however, and certainly where personality tests are the chief concern, it suffices to consider traits whose manifestations are not highly correlated.

b. Class models

When a clinician states that the meaning of a symptom, say, withdrawal from social situations, must be judged by the context in which it occurs, he implies a different structure from the cumulative one. Translation of this kind of clinical judgment into a structural model is not entirely obvious. One might translate it: of the several manifestations which characterize this trait, all, or perhaps all but one or two, must be present in order for one to say that the trait is present. The trait, then, is either present or absent. Its manifestations do not indicate amount, but if a sufficient number are present, signalize the presence of the trait. A smaller number have no significance in relation to the given trait. Clearcut clinical examples which follow this structure are not easy to find, whether because of the fluidity of present nosology or because clinicians are accustomed to thinking in more idiosyncratic terms. To take an example from medicine, fever and cough may indicate several diseases, fever and rash may also indicate several diseases, but fever, cough, and rash together strongly suggest a diagnosis of measles. There are other characteristic signs, particularly the history of onset, but some variation is observed. An example from test theory is Harrower-Erickson's [53] nine Rorschach signs. Presence of five or more is said to indicate neurosis. The example is not a perfect one, since the signs are all summaries rather than simple responses.

The foregoing pattern might be called a class pattern. It describes a class of people rather than a trait present to greater or less degree in all of us. Traits of the latter type, present in everyone to a degree, often follow the cumulative pattern described at the beginning of this section and might be expected to arise in relation to developmental experiences shared by all or by large segments of the population. Class patterns might be expected to be appropriate in relation to the problem of diagnosis. Historically, however, the patterns have not been used in that fashion. Hathaway and McKinley [54] used classification of patients into diagnostic categories as part of the original validational enterprise of the MMPI.

But the supposed traits defined by keys chosen for ability to discriminate diagnostic groups from normals were then treated as quantitative traits, as if they measured a characteristic more or less present in everyone. It may be true that various diagnostic groups represent extreme cases with respect to universal traits. Such a proposition requires proof, and the success of the original keys for the MMPI has not been such as to constitute such proof. According to French's [32] review, many of the personality traits found in factor analyses have been established either in normal populations or in pathological groups but not in both. (Construction of a model adequate to the complexities of clinical diagnosis will not be attempted here.)

Lazarsfeld's [65, 66] latent structure analysis is a general set of methods for handling data which do not conform to Guttman's scale model. In principle the methods are very general indeed; in practice the only methods which are well elaborated are those which account for a set of responses to items by assigning the individuals to (an arbitrary number of?) latent classes. Each latent class is homogeneous with respect to the trait presumed to be measured. The items are assumed to be pure measures of the trait, in the sense that all correlation between them is accounted for by that trait. Thus, within any latent class there is no correlation between items. In addition to the assumption about the number of latent classes, some assumption must be made about the form of the regression of items on latent classes. In general, both monotone and non-monotone items can be included (unlike most other methods), and the latent classes may or may not be assumed to be points on a single continuum. Observed frequencies of the various patterns of responses are then used to make inferences about the proportions of the population within the several latent classes. Those inferred proportions are utilized with the observed item frequencies to generate the expected frequencies of various response patterns. The expected frequencies are then compared with observed frequencies in order to test the goodness of fit of the inferred structure.

Hays and Borgatta [55] have shown that more than one assumed structure within the framework of latent class theory can generate data which agree reasonably well with a single set of observed data. This problem has been recognized by Lazarsfeld but has not been sufficiently investigated, at least in available work.

An investigation by Chiang [12], although not mentioning latent structure analysis, appears to be concerned with essentially the same mathematical problem. Chiang demonstrates that for four or five items the power of the test of any latent structure is poor unless the number of individuals is in the thousands. To have a reasonable chance of detecting the falsity of an assumed structure, even with about a thousand cases, eight items are needed. Most if not all published instances of latent structure analysis have utilized fewer than eight items; typically, four items are used.

In view of Chiang's work, or, to put the matter another way, in view of the many arbitrary assumptions which must be made to apply latent class analysis, there is some question how valuable a tool it will prove to be in terms of contribution to theory. Strangely enough, published illustrations of its use concern chiefly traits which are more easily conceptualized as quantitative than as categorical, such as soldiers' attitude towards the Army [64].

c. *Dynamic models*

Neither the cumulative models nor class models are adequate for dynamic traits. Psychoanalytic theory postulates, and some psychologists who reject other aspects of psychoanalysis also accept, that opposing tendencies may have a common motivational source. Peak [92] assumed that dynamic organization of traits could be demonstrated only by time changes. The view of the present monograph is that it

should be possible to find structural traces of dynamic organization, though admittedly the crucial evidence is not yet in.

One dynamic structural possibility is that two particular manifestations of the same trait may be mutually exclusive or in a less extreme case, negatively correlated. Cattell [11, Ch. 5], who discussed this possibility at length, showed that the effect of this structure may be to attenuate a positive relation. Frenkel-Brunswik [33] found in a study of adolescents that ratings of overt behavior on "exuberance" and "irritability" correlated .30 and .42 with intuitive ratings of the drive for aggression by clinicians who knew the *Ss* well. The two behavioral ratings were negatively correlated ($r = -.52$) and the multiple correlation of drive for aggression with exuberance and irritability was .73. She cited these data as evidence for the principle of alternative manifestations of a drive [33, p. 302].

A somewhat different dynamic structure may be derived from the notion of neurotic conflict. The notion that conflict is characteristic of neurosis and self-consistency characteristic of normality has been the basis of a number of researches, including those of Zubin [120], Winthrop [116], and McQuitty [83]. Winthrop [116] defined consistency in terms of formal logic, thereby ignoring a distinction made earlier by Zubin [120], among others, between the logical and the psychological meaning of an item. It is thus not surprising that his chief empirical discovery was that normal college students responded to his test in a markedly illogical fashion.

Zubin [120] defined consistency in terms of empirically determined patterns of response, differentiating, however, patterns characteristic of normal from those of pathological groups. His normal group had on the average the most patterns altogether and the fewest abnormal patterns. The neurotic group had on the average fewest patterns altogether despite having the greatest number of abnormal patterns, exceeding in both respects the psychotic and the organic groups of patients. Zubin's method, which consisted essentially of looking for all possible patterns of response, is unwieldy because of the large number of possible patterns with a reasonably large pool of items. He solved this problem by examining only a small fraction of the possibilities, clearly an unsatisfactory solution.

Several studies by McQuitty [83, 84] have also been concerned with measurement of "personality integration" and with testing various hypotheses about the relation of such integration to mental health. While considerable ingenuity in the invention of methodology was exercised, the methods are not comparable with those discussed in the present monograph.

An unpublished study has taken the same basic idea, that conflict is evidence of maladjustment or neurosis, and combined it with content considerations. Dr. Jack Block,[3] in investigating the postulated dimension of ego-control, found the most maladjusted individuals tended not to be those extreme in the direction of over-control or extreme in the direction of under-control. Rather, the psychologically most disturbed individuals appeared to be those who manifested, in some extreme, inconsistent, unintegrated fashion, both over- and under-control simultaneously. This pattern differs from the one of alternative manifestations, since there is no necessary contradiction between one particular manifestation of control and another particular one. The contradiction resides in the individuals, and the techniques for demonstrating the existence of the contradiction must be correspondingly different.

From a formal point of view the problem of "scatter" in measurement of abilities resembles the problem of dynamic structure. Two quite different kinds of scatter have been studied, intra-test and inter-test scatter. In the case of intra-test scatter, the

[3] Mimeographed report, "The development of a MMPI-based scale to measure ego-control," issued at the Institute for Personality Assessment and Research, University of California, 1953.

test items are all assumed to be equivalent in function and the dispersion of successes and failures along the scale of difficulty (or its analog) represents S's tendency to inconsistency. In the case of inter-test scatter, the functions called on by the several tests are not assumed to be identical. Many studies have attempted to show, with varying degrees of success, characteristic patterns of relatively high and low scores for different clinical syndromes [39, 59]. The Wechsler-Bellevue test has been most often studied in this regard, but other tests have also been used. Even when studying inter-test scatter there must be an assumption of community of function in the several subtests, else there would be no basis for establishing a general level and thus of measuring scatter. The general level is established by some aspect of S's performance, often by a test such as vocabulary which is presumably not easily subject to deterioration. The two types of scatter might be termed "pure scatter" (intra-functional) and "patterned scatter" (inter-functional). The possibility must be recognized that what appears to be pure scatter may in fact be an unrecognized pattern. Both types of scatter have theoretical implications, but patterned scatter can be expected to be more fruitful theoretically. There are many possible ways of measuring scatter; theoretical implications of the various coefficients have not been fully explored.

3. *Pattern analysis and configural scoring*

The topics of configural scoring and pattern analysis [35] have received attention in many quarters recently and must be related to the present discussion. All of these methods, with the notable exception of a recent exposition by Haggard [52], appear to be criterion-oriented; that is, they are methods of combining data so as to improve prediction of a single specified criterion. Cronbach and Gleser [19] have shown that measures of profile similarity in general are simple measures of linear distance in a space equal in dimensionality to the number of tests; such models assume linear or additive combination of tests. The term "configural scoring" has been used for non-linear combination of test and item scores [86, 56]. Methods such as Lykken's [80] actuarial pattern analysis and duMas' [24, 25] manifest structure analysis include both linear and non-linear models.

Meehl [86] introduces his discussion of configural scoring with a paradox: he displays an instance where each of two variables fails to predict the criterion but the two scored together have high validity. But here Meehl has introduced a paradox vaster than he apparently intended, and it is the paradox of all actuarial pattern analysis. The broader paradox is this: no item of information can ever be discarded as useless for any prediction. For even if it seems to have no correlation with the criterion, it may, when scored configurally with other items, yield valid prediction. There is no means for assigning any reasonable limit to the number of items which must be considered as possible predictors, nor to the possible combinations which must be examined. Shall we measure the universe and intercorrelate everything in it? Worse, there is no limit to the variety of relationships which must be explored.

In practice, suggestions of various kinds are made for limiting the scope of the search. Indeed, in view of the present access to electronic computing machines, the limiting consideration is not the examination of large numbers of combinations of data but the fractionation of the sample to the point that the number of cases for most patterns is too small to be useful. Collection of patterns which exceed the level of significance only by chance would also appear to be a problem where almost unlimited instances are available.

Lykken [80] proposes to collect criterion data for MMPI profiles by using 9 scales. If each scale is dichotomized, as he proposes, there are still 512 possible patterns. One wonders how Lykken is confident that the non-linear relationships which actuarial

pattern analysis is to reveal do not occur within the dichotomized segment of a scale rather than between segments.

One difference between the approach of the present monograph and that of the various advocates of pattern analysis is that the latter do not separate clearly the steps of assembling a pool of items and of selecting the best items from the pool. Orientation of pattern analysis towards prediction of specific criteria rather than towards measurement of traits is another difference. Related to both of the foregoing points is the differing utilization of theory in pattern analysis and the present approach.

Tukey [111] has criticized a symposium on "Statistics for the clinician" for the paucity of theoretical considerations, the tendency of many psychologists to throw the whole problem of finding relationships into the lap of the statistician rather than providing hypotheses to guide the statistics. This approach has not been the pathway to advances in other fields, he indicates.

Tukey's point may be stated differently. We cannot cling to every bit of our data and still do scientific research, for every event is ultimately unique. For example, we agree to regard as identical all responses of "yes" to the question "Were you born in the United States?" regardless of the angle at which the pen met the paper, etc. Some generalization has already taken place when, say, all responses to a dichotomous item are classed in one of two categories. This classification is justified by our beliefs about the greater relevance of some aspects of response than others. Two different ways of saying yes are considered equivalent. This matter is discussed in greater detail by Meehl [87, Ch. 6].

For any considerable number of items to be manageable, there must again be a decision that certain groups or patterns of response are equivalent; in principle, this decision does no more violence to the data than the one already taken. It is at this point that I would apply Tukey's suggestion that we utilize theoretical (or at least substantive) considerations. The choice of a structural or statistical model, here advocated to be based on psychological considerations, automatically reduces drastically the number of response patterns which are considered different. For example, with a cumulative model scoring plus on any three keyed items is equivalent to scoring plus on any other three keyed items. The net effect of ultra-empiricism is that data reduction, choice of intervening variables, takes place in the dark—"dichotomize every MMPI scale at the median"—instead of in the light of theory.

4. *Use of structural models*

In view of the relative novelty of the concept of structural validity, an exhaustive list of possible structural relations prevailing among the several manifestations of a single trait is not in order. The purpose of this section chiefly is to point out that psychological theories postulate a variety of such structures. The kinds of structure explored above have been classified as cumulative models, class models, and dynamic models. While cumulative models differentiate individuals with respect to degree and class models differentiate them with respect to kind, use of scatter in diagnosis implies that either the degree of structure (intra-functional scatter) or the nature of the structure (inter-functional scatter) is itself what differs among individuals.

Much of what has been called structural theory in contemporary psychometrics has not been included in the foregoing discussion. The reason lies in another terminological misfortune: at least two different kinds of structure are involved. That to which the present paper refers is the structure which subsists among the various manifestations of a trait. A name for this kind of analysis might be *score-structure analysis*.

Most of the work of Guttman and Coombs and in the field of factor analysis

has been thought of as analysis of the components which contribute to the causation or formation of tests responses or scores. This kind of analysis may be called *component structure analysis.* (Lazarsfeld's latent structure analysis is similar to score-structure analysis.) While such analyses have been conceived in terms of components, the data have, of course, been the same as for score-structure analysis, and the formal models and their properties may prove to have application in a different manner than their originators intended. Guttman's [50] simplex theory, an aspect of radex theory, which is a kind of generalization of factor analysis, has been explicitly worked out both in terms of component analysis and of score-structure analysis.

To analyze the manner in which the component traits determine responses and scores is a more fundamental enterprise than simply studying which responses go together in some fashion. The former pursuit is beset with difficulties. The structural component of validity, as conceived herein, is not concerned in the first instance with the mediational pathway of the observed responses. Such mediational pathways are part of the "theory" which it is hoped that objective tests will illuminate. The manner of achieving that illumination is the subject matter of the present monograph and its predecessors [20, 92, 60]. Such mediational mechanisms as psychological theories have postulated are often far more complex than can be accommodated by present component models. Such mechanisms as condensation, displacement, isolation, and reaction formation elude not only additive models but also the more complex models of Coombs (compensatory, conjunctive, disjunctive) [15] and Guttman (simplex, circumplex, radex) [50, 51]. With further development of the theory of objective test behavior and further development of component models, the latter may better serve the former.

It seems unlikely, though, that the path from data to theory will be greatly illuminated by single mathematical coups, however brilliant. Multiple factor analysis has an enormous advantage over recently proposed models. The conditions for making inferences from data, as formulated, e.g., in a recent paper on simple structure by Tucker [110], are the outcome of mutual criticism of many workers over a long period of time and in connection with a wide variety of substantive problems. Work of Guttman, Lazarsfeld, and Coombs is accompanied by no such clear-cut rules for inference; indeed, escape clauses are frequent, such as, "Sampling errors will not be considered."

The reader is urged to keep the following questions in mind in utilizing or evaluating the use of structural models:

- Does the chosen structure conform to what is known of non-test manifestations of the putative trait?
- Is the degree of structure or degree of conformity to the model quantitatively evaluated?
- Does the model impose structural characteristics on the data? To the extent this is true, no theoretical conclusions can be drawn from the structure.
- Is the model used for selecting data? If so, degree of structure should be evaluated on a new sample not so used.
- Are the parameters of structure (number of factors, number of classes, etc.) uniquely determined? To the extent that they are not, caution must be exercised in ascribing theoretical importance to the results.

Proponents of various structural models sometimes study narrow or even artificial kinds of data, then indicate with a wave of the hand that all other kinds of data will also be illuminated by the given model. Periodic reviews of structural theory should encourage wider and more discriminating use of available models. Nor will the light shine in only one direction. One can only deplore the presentation of psychometrics in terms of hypothetical monsters whose test answers are determined by the spinning of internal roulette wheels [66].

The surest beacon for psychometric theory is real involvement in the solution of psychological problems.

C. EXTERNAL COMPONENT

While the problems of substantive and structural criteria for the construct validity of a test were treated sketchily in the *Technical Recommendations* and by Cronbach and Meehl and have consequently been treated at length above, the problem of external criteria for validity was ably treated. Many aspects of the problem discussed recently will not be repeated or reviewed.

In discussing substantive and structural components concern has been with responses to individual items. While individual items may be studied in relation to external criteria, most of what comes under the heading of external validity concerns correlation with total score. The method of constructing a total score from the item pattern necessarily implies a commitment about the structure of the items, and thus about the structure of the trait measured. That is, in a cumulative test, where the total score is the number of items scored plus, an additive model is implied. The indiscriminate use of this model may account for some of the difficulties which psychologists encounter in attempting to measure dynamic traits. The total score, with its implied commitment as to structure, is not a datum of observation but is an intervening variable.

External validity may be subdivided in various ways. The *Technical Recommendations* use a temporal division, concurrent and predictive validities. This division appears to be almost arbitrary and inconsequential. But determining whether right now X is psychotic is different from predicting whether at some future time X will become psychotic. The terms *discriminative validity* and *predictive validity* convey the distinction.

Another division of external validity compares the relation to other test scores with the relation to non-test behavior and other non-test data. By relating the test scores to other tests, one can make indirect use of whatever is known of the validity of the other tests. Factorial pattern comes under this heading. Surely no test, however, is acceptable for clinical use unless it is in some way tied to a non-test criterion, whether that be behavior ratings, group differences, or whatever. The warning of the *Technical Recommendations* [121, p. 15] is cogent here: the test must show relevance to non-test behavior, but it need not be equivalent to non-test behavior.

Another subdivision of external criteria is into those which are predicted to show a positive (or negative) relation to the test score and those which are predicted to show no relation to the test score. The problem of variables which ought not to show relation to the given test score is that of distortions of measurement. Distortions are errors of measurement which are correlated with true scores or with other obtained scores rather than being random with respect to all other scores as classical test theory assumes. An example of a distortion is a response set [17]. A similar example is the IBM answer-sheet-marking factor found by Whitcomb and Travers (personal communication) to overshadow other effects which they wished to study. Facade can act as a distorting factor in studies which are intended to measure other personality traits; since it is found so ubiquitously, it should be studied more thoroughly in its own right. All of the factors responsible for secular trends in test responses, to be discussed under Secular Trends in Test Behavior, are also distortions, as are unfinished items in non-speed tests. Demonstration of negligible relationship with known sources of distortion is an essential, not an optional, step in test validation.

Cronbach and Meehl [20] discussed correction for distortions of measurement in terms of suppressor variables. Recently Brogden [6] has proposed a method of constructing forced choice items which

makes use of the principle of the suppressor variable. The objection to suppressor variables is, of course, that they introduce another fallible score and thus another source of error variance. Loevinger [72] has proposed an alternative approach of selecting originally those items least weighted with distortions.

Eysenck [28, 29] has proposed a method, called "criterion analysis,"[4] which combines external validation and study of internal structure. Briefly, he introduces into a factor analytic study a measure which represents the criterion. The tests are factored by any of several methods, and the axes rotated so that one axis coincides with the criterion measure. His method is therefore an alternative to simple structure as a principle for rotation of axes. Sidestepping the technical problems, some of which are discussed by Lubin [79], the question arises how the method of criterion analysis relates to the two steps of ascertaining structural unities and subsequently analyzing external correlations. Is the same thing accomplished, something more, or something less?

One may note, first, that complete dependence on factor analysis restricts one to an additive structural model, but this is perhaps a minor point. In some examples of criterion analysis quoted by Eysenck there is clear evidence of significant distortion in the criterion. For example, groups which were intended to be matched except for the fact that one group was rated as neurotic and the other was a normal control turned out to be significantly different in age and intelligence. Rotation towards such an axis produces a melange of neuroticism, age, intelligence, and who knows what else. Eysenck fails to recognize that such flaws in criterion groups and corresponding distortions in criterion measures are the rule rather than the exception. In many problems control groups are almost a contradiction in terms, or, as Rogers, *et al.* [94], found, somewhat destructive of the purpose of the study. Such difficulties in external criteria led to the extension and revision of the concept of validity in the direction of construct validity.

The regularity with which measures of motor control and speed turn out to be the most promising measures of neurosis in studies under Eysenck's direction induces the suspicion that age and/or general physical condition is the chief component of his "neurotic tendency" factor. In one major study [29, Ch. 4] a measure of intelligence was more highly correlated with the neurotic-normal criterion than any other measure except one questionnaire, a fact somewhat concealed by his *ad hoc* explanations. Neither criterion analysis nor any other computational gimmick takes the place of careful consideration of possible sources of distortion.

While Eysenck's method of criterion analysis seems to be a step backwards towards criterion-oriented validity, it does point to a weakness in attempts to establish the existence of traits by means of factor analysis alone. It seems reasonable to require that complete validation of any test include a demonstration of some non-zero relationship with a non-test variable. Cronbach and Meehl specified that validity must include some external relationships. They did not specify non-test variables, but the illustrations adduced in their paper appear to indicate sympathy with such a requirement.

In Section B of this chapter, choice of structural model was presented as an intuitive problem. Magnitude of external correlations may help to decide between alternative structures. In *Measurement and Prediction* [97] there is record of attempts to find scalable aspects of neurotic tendency. A clinician might have told the investigators that neurosis is far too variable in its manifestations to conform to a scale model, but apparently none did. There were several sets of items which proved to be scalable, but such items were extremely

[4] The term "criterion analysis" is sometimes used to cover a broad range of techniques, rather than just the one so designated by Eysenck. See, for example, Lafitte [63].

poor at differentiating neurotic from normal soldiers. The one subtest (psychosomatic complaints) which was more or less adequate at differentiating neurotics was not scalable but formed what Guttman calls a quasi scale. Very likely such items would best be handled by classical test methods, that is, a cumulative model assuming low intercorrelations between items. Even where the structure of the items is of central interest, external correlations provide the court of last appeal.

SECULAR TRENDS IN TEST BEHAVIOR

What has been presented in this paper so far may be called a *theory of the first test.* It presents in rough outline a complete theory for construction and evaluation of a test without reference to test-retest coefficients or parallel forms of tests. Part of the traditional theory of reliability, the part that deals with homogeneity, has been assimilated to the topic of the structural component of validity. The other part of reliability theory, the part that deals with stability of test scores, has been omitted and does not appear to be needed. One cannot, however, propose a test theory which bypasses the major part of traditional theory without serious presentation of reasons.

A. CIRCULARITY OF CLASSICAL RELIABILITY THEORY

Consider the statistical theory of reliability. On a previous occasion [69, Ch. 1] I have argued that reliability can be defined only in a circular manner. Gulliksen's [43] definitive summary of test theory, which has become available since the argument was first made, illustrates it well. Gulliksen derives the basic equations twice, first in terms of a definition of random error, then in terms of a definition of true score. Both derivations begin, of course, with the assumption or definition that the obtained score is the sum of true and error scores.

Under the heading, "Definition of random errors," the error score is defined as having a mean of zero, zero correlation with true score, and "*the correlation between errors on one test and those on another parallel test is zero*" [43, p. 7, italics in original]. Under the heading, "Definition of parallel tests in terms of true score and error variance," Gulliksen states, "*For two parallel tests, the errors of measurement are equal*" [43, p. 12, italics in original]. Thus parallel tests are needed as part of the definition of error scores, and errors are needed as part of the definition of parallel tests.

In the following chapter, derivations are in terms of true score. True score is defined as the limit of the average score on a number of parallel tests, as the number increases without limit. Parallel tests are defined as tests equal in observed means, standard deviations, and mutual intercorrelations [43, pp. 28–29]. Reliability is then defined as the correlation between parallel tests [43, p. 31, footnote]. Thus parallel tests are defined in terms of their correlation (reliability), and reliability is defined as the correlation between parallel tests. It is not surprising that a theory which begins with a circularity ends in paradox [71].

The basic difficulty in the theory of reliability can be stated other ways. The true score is defined as the limit of a series which may not in fact converge, or, if it does converge, is not approximated usefully by the first term in the series, which is the term whose meaning is at stake. Or, reliability is defined in terms of a set of operations, to wit, repeated testing with no effect of repetition, which cannot in the nature of things be performed.

To circumvent or ameliorate this basic contradiction various methods of estimating reliability have been employed. There has for a long time been some recognition and recently widespread recognition that estimates of reliability based on a single administration of a test are estimates of something completely different from what is estimated by a test-retest coefficient. That

the same statistical theory of reliability should continue to be used for both kinds of coefficients is remarkable testimony to the obduracy of psychometricians. Correlation of parallel forms of tests combines some of the advantages and some of the disadvantages of test-retest and single-form reliabilities and, in addition, involves the circularity of parallel forms being defined in terms of the very correlation which they will then be used to estimate.

Gulliksen says of the parallel forms method of determining reliability: "Generally speaking, this method is best, provided that we can regulate the interval between the two tests and the activity of the subjects during that interval so that the influence of practice, fatigue, and other similar effects will be negligible" [43, pp. 214–215]. The question arises: for what kinds of psychological tests has it been established that influences of practice, fatigue, boredom, sophistication, and other temporary or cumulative influences have a negligible effect on the second test? It is probably true, as Gulliksen claims, that such effects on test behavior, which will here be called *secular trends*, can be detected in terms of change either in mean or in variance. But traditional test theory assumes that secular trends in test behavior do not exist, and most reliability coefficients are reported without reference to whether they have or have not occurred.

To belittle the advance which reliability theory represented over the earlier naive view of test scores would be both folly and impropriety. The recognition that the obtained test score was an imperfect estimate of something more fundamental was a tremendous step forward. The various coefficients for estimating reliability and the statistical theory of reliability arose in attempting to assay how well obtained scores succeeded in measuring the more fundamental or "true scores." The import of the preceding chapters of this monograph is that recent developments in test theory, particularly the concept of construct validity, enable us to conceptualize what the obtained score is an estimate of and to evaluate how good an estimate it is without reference to a retest or a parallel test.

The basic assumption of reliability theory, which implies that secular trends do not exist, has been carried over into a number of recent contributions to psychometrics. Guttman [45] has written papers on estimation of test-retest reliability coefficients; however, the assumption that there is no effect from repeating a test does not appear to be incorporated into his theories of scale analysis, of image analysis, or into radex theory. Coombs and Lazarsfeld, however, have incorporated the classical assumption into their thinking more essentially. Lazarsfeld [66] uses the term "brainwashing" to help his readers conceptualize what is assumed to take place between test and retest. Anderson [3], however, has presented a computational technique for latent structure analysis which does not require estimation of item reliabilities.

B. THE PROBLEM OF SECULAR TRENDS

Consider this possibility. A test of ability given without forepractice proves to have high predictive validity, but test-retest correlation is low. Given with considerable forepractice the test-retest coefficient is high but the nature of the test has changed so that predictive validity is low. Here is another "attenuation paradox" [71], i.e., a set of experimental conditions under which raising reliability lowers validity, contrary to the relationship stated in the classical correction for attenuation. The example is by no means implausible. It was proposed by Thorndike [102] in 1919. It demonstrates the intimate relation between secular trends and the validity problem.

The necessity of conforming as far as possible to the assumptions of reliability theory has led to a neglect of and glossing over of the problem of secular trends in test behavior. There is every reason to suppose that in general they occur. The "theory of the first test" can ignore them. But in practical and also in research situations retests

and parallel forms are often necessary. Second tests, whether retests or parallel forms, must be interpreted with respect to most or all of the factors bearing on first tests and also all of the factors influencing trends in test behavior as such. Thus coefficients of stability, when properly identified as to the conditions under which they were obtained, have an important place in test theory. They belong, however, to "theory of the second test." On occasion results from further testings will also be available, as in longitudinal studies.

One reason for the importance of the topic of secular trends is that such trends create artifacts, either spurious stability or spurious change, in research primarily directed toward study of stability and change in other traits. Thus the topic is intimately related to the subjects of trait consistency, developmental changes, intra-individual differences, and pathological and therapeutic processes.

A major conceptualization and review of the literature on intra-individual response variation has been made by Fiske and Rice [30]. They separate intra-individual response variation into three types. Type I consists of variation in response due to spontaneous change in the organism and is assumed not to depend on the order of presentation of stimuli. Type II includes variation which may be influenced by the order of presentation, but effects of fatigue, learning, and cyclic change are arbitrarily excluded. Type III includes changes where the stimulus situation changes, as in studies of rigidity, and is included to throw the previous types into perspective. Insofar as variation in response is spontaneous and not a function of order of presentation (Type I), the assumptions of traditional reliability theory would appear to be supported. Fiske and Rice, however, believe that such variability is lawful and has correlates, and that belief appears to be somewhat at variance with the traditional assumption of random error; the contradiction is not direct, for the traditional assumption is concerned with a quantity, Fiske and Rice with its standard deviation. It is not clear how rigorous were the standards for deciding whether a given study exhibited Type I or Type II variation. (Indeed, the classification of a study, or even the existence of the defined types of variation, represents a complex inference. It would appear generally desirable for review articles to adopt less disputable categorizations of studies.)

Type II variation corresponds closely to what is here called secular trends. However, systematic factors such as fatigue and learning are not excluded from secular trends insofar as they produce distortion in measures of other traits. Despite the heroic proportions of the Fiske and Rice review, which had 233 references, important evidence for the existence of secular trends was omitted, such as evidence for rise in IQ with repeated testing, even when the tests are as much as a year apart.

Among early papers on systematic changes on retest for ability are those of Thorndike [102] and Adkins [1]; neither is mentioned by Fiske and Rice. Windle [115] has reviewed the literature on "Test-retest effect on personality questionnaires." He presents a table of 41 studies of retest of untreated groups; of those 41 studies only 4 are listed in the bibliography of Fiske and Rice's paper. The general trend of the studies reviewed by Windle is for retests on personality inventories to show improved adjustment.

Bayley [4] in a recent review has shown that increments in measured intelligence where the same individuals are retested exceed previous expectations, even when the retest occurs after a considerable time lapse. She acknowledges that practice effects probably account for some of the increment but lays greater stress on the possibility that intelligence continues to develop during maturity.

Recognition that systematic changes occur on retest has coexisted with classical reliability theory over a period of many years, with only rare recognition of the contradiction between fact and the assump-

tions of theory. Clearly there are questions here which need clarification by further evidence or by re-examination of existing data. The concept of secular trends in test behavior is needed to prevent dependence on traditional test-retest correlations from obscuring substantial problems, those of changes in means and variances.

C. IMPLICATIONS FOR CONSTRUCT VALIDITY

That there is no received method for dealing with secular trends is borne out by inspection of the *Technical Recommendations*. Secular trends are mentioned just once, in connection with stability of scores: "The manual should report changes in mean score as well as the correlation between the two sets of scores" [121, p. 232]. That requirement is considered essential, but there is no follow-up. What should be done in case there is a sizeable mean change from test to retest?

The common belief, explicit in Peak and not disavowed in the *Technical Recommendations* or Cronbach and Meehl, is that knowledge of stability contributes to our confidence in the construct validity of the test because we ordinarily are interested in traits which have some longevity; stability should not exceed, however, whatever stability might reasonably characterize the underlying trait. The traditional correction for attenuation is often interpreted as a precise statement of relation between stability and validity; however, it belongs to the part of psychometric theory which treats coefficients of homogeneity (such as split-half correlation) and coefficients of stability (such as test-retest correlation) as identical.

The supposed connection between test-retest (or parallel forms) stability and construct validity is, I submit, a *non sequitur*. The basis for the *non sequitur* is failure to recognize the fact that the first and the second tests may, and in general must be assumed to, bear different relations to the underlying trait, just by virtue of being first and second. This topic is another aspect of the "psychology of objective test behavior." In the case of projective tests, secular trends in test behavior are so conspicuous that the notion of test-retest stability often is not applied to them [16]. Yet secular trends of a less conspicuous sort are regularly found in other kinds of tests; the difference is one of degree.

Suppose a mean gain of 5 raw score points is found on retesting children with a certain test in a given age range. What procedure should the test user follow? Should he subtract five points from each second test and then refer to norms for the first test? The *Technical Recommendations* do not say. So doing would ignore any possible change in variance on retest. Separate norms for retests would seem more defensible. Failure to recommend separate norms for retests is perhaps the only serious criticism that can be made of the work of the APA Committee on Test Standards.

Perhaps, however, the implication of the *Technical Recommendations* is that if the change in mean score from test to retest is appreciable, the validity of the test is open to question. But if the correlation between test and retest is much less than unity, validity of test and retest may differ considerably.

Assume the validity of the first test is investigated first. Neither change in mean score nor low correlation between test and retest places any restriction on the validity of the first test. If the test-retest coefficient is very high, then the second test may be assumed to have similar validity to the first, provided care is taken to use retest norms. Since high correlation is consistent with sizeable changes in mean and standard deviation, referral of retests to first test norms may seriously impair the validity of resultant inferences.

Occurrence of an appreciable mean change does not in itself imply any decrement in test-retest correlation. Intuitively, however, one assumes that some factor

must cause the mean change, and that this implies a factor influencing the retest which does not influence the original score. On this reasoning one would be surprised if a sizeable mean change would occur where test-retest correlation was near to unity. Such coefficients as are commonly found do not rule out the possibility of fairly large discrepancies between correlations of the two tests with external criteria and other variables.

In conclusion, neither the magnitude of the test-retest coefficient nor the magnitude of changes in mean and variance from test to retest has direct bearing on the construct validity of the first test. Norms for the first test can properly be used in the interpretation of retest scores only when the correlation between the two tests is very high and changes in mean and standard deviation are negligible. In other cases the retest is, from the point of view of construct validity, a new test, and must be validated and, especially, standardized accordingly.

ALTERNATIVE APPROACHES

The title of the present paper outlines a province that many psychologists will recognize as their own. The scope and length of this monograph and, to be candid, the cathexis of my own view preclude a thorough review here of alternative ways of organizing the field. At the same time it would be misleading not to acknowledge major predecessors. The reader will recognize that the intention of this chapter is clarification of the present contribution by contrast rather than a disinterested exposition and evaluation of alternatives.

A. OTHER KINDS OF DATA

Approaches utilizing rating scales and projective techniques have been excluded from consideration but it seems desirable to relate such techniques briefly to the discussion. Consider first rating scales. As they are ordinarily used, the trait rated is identical with the trait in which one is interested, and the problem of substantive validity cannot arise. Similarly, the question of structural validity cannot arise in the case of a single rating of a trait, since the rater is expected to weigh and evaluate all the manifestations of the trait before making his rating. Thus the sole criterion for this kind of rating is external validity.

Rating scales can be used in a manner formally similar to test items. Wittenborn [118] has utilized ratings of specific symptomatic behaviors in a factor analytic study. The resultant clusters have been examined for substantive and structural validity and do in fact appear to make psychiatric sense.

Suppose there were an objective test with 5000 items, and each *S* is instructed to answer "some" of them. The number of answers per *S* might range, say, from 5 to 500, and of course different *Ss* answering the same number would generally choose different items. The question of which items are interrelated can be asked with meaning. But unfortunately the problems of how many responses different *Ss* make and which questions they choose to answer, while themselves meaningful data, make it virtually impossible to study the relations of the items. The illustration is analogous to projective tests, except that in the latter the number of possible responses is far greater. The price paid for the richness of the response as a reflection of the mental processes of the S is the difficulty in comparing data from several *Ss*.

Any kind of score for a projective test, like a score for an objective test, is an intervening variable rather than an original datum. The response to an item is a datum, a single manifestation of personality or item of behavior. Structural validity refers to the structural relations of the original data and not of the intervening variables or scores. Evidence for the structure of the *items* is what establishes that a *scorable test* exists in the data. Thus the structural component of validity, as the term is used here,

although not meaningless for projective tests, is difficult to establish. Wittenborn [117], again, has devised some approaches to the problem of the structural validity of certain Rorschach scoring categories. Whether he has satisfactorily solved the difficult methodological problems is not entirely clear. The substantive component of validity, although used in relation to objective tests as referring to items, appears to be meaningful when applied to interpretation of Rorschach scoring categories.

Schafer [95], on occasion, uses the sequence of responses in a Rorschach administration as a map of the personality dynamics of S. The rules for such inference are not at present sufficiently explicit so that group differences can be clearly predicted. Empirical verification of the rules is still not planned for. This kind of interpretation can be compared with the notion of structure as used in the present paper. The suggestion here is that there are populations with respect to which one has certain intuitions about the structural relations of the several manifestations of one trait. This structure should also obtain in the several items which comprise a test of that trait. The same two assumptions appear to be made by Schafer, but his technique skips the step of finding those items which best conform to the given structure and instead utilizes structural (sequential) differences as measures of trait differences. Any non-equivalence in the stimulus situation preceding the several responses must be taken into account intuitively.

Interpretation of response sequence may also be compared to the notion of secular trends in test behavior. Again there is an assumption of the equivalence of the several stimulus situations. Except for this instance, there are not many examples in psychological testing of constructive use of factors making for secular trends in test responses.

A recent study by Wittenborn, *et al.* [119] is of interest both because it utilizes data from interviews in a manner formally similar to objective test data and because Wittenborn has evolved a method superficially similar to that advocated in the present monograph. The differences between Wittenborn's method and that of the writer should sharpen the focus of the present argument. The purpose of Wittenborn's monograph is to demonstrate that interviews yield data from which scores can be obtained which are usable in the same manner as ordinary test scores. He speaks of standard "bits of information" derived from interview protocols. The phraseology suggests the notion of items as observations as opposed to miniature measurements, as developed under Relation of Test Behavior to Theory, Section C, above. Wittenborn, however, studies the interrelations of such items by means of tetrachoric correlation, a coefficient which assumes not only that each item is actually a measurement of an underlying continuum, but also that the continuum is normally distributed, the correlation between any two items is characterized by linear regression, and so on.

Further, Wittenborn searches for clusters among the items of information scored from the interviews. Unfortunately, he makes no use of statistical methods worked out for this problem, such as the method of Wherry and Winer [114] or that of Loevinger, Gleser, and DuBois [75]. The latter method, in particular, provides formal solutions for two problems: when to stop adding items to a cluster and when two clusters should be combined. Wittenborn does not appear to have satisfactory solutions to these problems. For example, he presents two separate clusters whose intercorrelation is .56 but whose self-correlation is in each case .44. Although the split-half correlation reported is less appropriate as an index of homogeneity than one based on the Kuder-Richardson Formula 20, they are probably not far apart in magnitude. These data strongly suggest that the between-cluster item intercorrelations exceed the within-cluster intercorrelations, which surely is contrary to the author's intention.

t data and because
d a method super-
: advocated in the
'he differences be-
thod and that of the
he focus of the pres-
ɔose of Wittenborn's
ɔnstrate that inter-
which scores can be
ɪsable in the same
st scores. He speaks
nformation" derived
ɔls. The phraseology
items as observations
re measurements, as
ɪion of Test Behavior
above. Wittenborn,
nterrelations of such
achoric correlation, a
ɪes not only that each
urement of an under-
: also that the con-
ɜtributed, the correla-
items is characterized
nd so on.
ɪ searches for clusters
ormation scored from
unately, he makes no
hods worked out for
'he method of Wherry
: that of Loevinger,
s [75]. The latter
provides formal solu-
ɪs: when to stop add-
r and when two clus-
ned. Wittenborn does
ɪtisfactory solutions to
example, he presents
s whose intercorrela-
e self-correlation is in
ɪgh the split-half cor-
less appropriate as an
y than one based on
n Formula 20, they are
rt in magnitude. These
st that the between-
ɔrrelations exceed the
rrelations, which surely
ɪthor's intention.

Finally, Wittenborn admits that content was taken into account along with coefficients in formation of clusters. Now if the intention of his monograph had been to provide scoring keys for clinical use, this procedure would be excusable at least as an interim measure. But the purpose of the monograph was to demonstrate that scorable clusters exist. For research purposes it seems imperative to separate the decisions which are guided by content from the decisions which are guided by numerical data. One method of doing so is outlined under Components of Construct Validity, above. Clusters are formed on the basis of data alone, without prejudice from content. Examination of content then serves as a check on the value and the interpretation of clusters or scoring keys thus derived.

B. OTHER KINDS OF ANALYSIS

The Berkeley studies on the authoritarian personality [2] exemplify the explicit utilization of theory in construction of objective test items. While the tests, F scale, E scale, and others, were modified as a result of empirical findings, it is not clear how much the test results yielded for theory in the Berkeley studies. Structural considerations were excluded in test construction. Likert-type items were used, introducing a spurious element into the intercorrelation of their several scales.

By far the chief instrument for using empirical relations of test behaviors to define constructs has been factor analysis. The many contributions of Thurstone are preeminent; perhaps in this country insufficient credit is given Spearman, Burt, and Thomson. Vernon [112], in summarizing factorial studies of ability, found support for the English predilection for hierarchical structure of mental abilities. More recently Guilford [41] has derived from a review of factorial studies a new picture of mental organization. Other investigators, too, have been concerned with the implications for psychology of the results of factor analysis. There is, however, some question whether factor analytic studies have entirely lived up to optimistic advance notices of their contributions to psychological theory. Alternative structural theories have been presented with equally optimistic advance notices, have less clear rules for making theoretical inferences, and have as yet yielded less for theory.

Suppose, as in the Q-sort technique, an S is given 100 first person statements, which he is to sort into piles according to how closely they characterize him. Does the distribution he arrives at differ from the distribution his neighbor arrives at in content or in structure? Neither notion seems clearly to apply. It is difficult to see what substitutes in the Q-sort technique for the demonstration in other objective tests that structure does exist, which implies not only a common trait underlying several responses but also a community of meaning in the reading of single items and test instructions. Relatively little use is made of responses in Q-technique, however. What is usually used is again an intervening variable, namely, the correlation between pairs of distributions obtained under different circumstances. These correlations are themselves treated as data and related to other variables, so that a hierarchy of intervening variables is established [94]. Relationship between the hypothesis tested and the data observed has then become tenuous indeed.

Finally, the reader will recognize that the approach of this paper has been cross-sectional and probabilistic-functional in Brunswik's [8] sense, or confined to "R-technique," if we may disregard for the moment Cattell's wish to confine the term to factor analytic studies. What of the alternatives developed by Cattell [11], Stephenson [98], and Mowrer [90], to wit, P-, Q-, O-technique, and so on? These techniques substitute tests or occasions as the dimension of replication for the more usual replication of people; in place of correlating tests, people or occasions may be correlated. The "covariation chart" in terms of

which the alternative techniques are developed is by itself, I believe, somewhat obstructive of progress. The feeling that it induces in some scientists, that any ways of collecting and collating data that are possible are also necessary, seems misguided. It is too much like the research plan of the mythical chemist: take substances from the storeroom in alphabetical order and place in the cyclotron. Where P-technique and Q-technique methods grow out of the nature of a problem investigated, they may well be the most appropriate methods, but that must be decided in specific context. As programs, their disadvantages seem to outweigh the advantages. The excessive time requirements of P-technique factor analysis will surely drive psychologists farther away from the generality of the population and into the company of special types of people. The basis for this conclusion is easily demonstrated by attempting to get the cooperation of various kinds of people for brief paper-and-pencil tests.

As the discussion of the structural component under Components of Construct Validity illustrates, the study of types and configurations is not the exclusive province of Q-technique, and the study of dynamic processes is not the exclusive province of P-technique. Many of the problems discussed above, such as the relation of secular trends to the validity problem, assume a different form in relation to alternative techniques.

Anyone who proposes a radical departure from traditional psychometric techniques must be prepared to assume the burden of proving the superiority of his methods to classical ones, for classical *methods* have clearly established their worth, at least in the field of ability measurement. But there is a lack of any coherent correspondence between classical psychometric theory and psychometric techniques; so pragmatic success does not support equally classical *theory*. The present monograph proposes a test theory which differs radically from the generally accepted one but advocates methods which are not widely at variance with classical methods. Considerations adduced in this essay do not, however, justify dogmatic rejection of more radical innovations of technique.

A PSYCHOMETRICS

A. THEORY

In summary, historical and litigious passages may be neglected in order to review what has been presented as an outline for a psychometrics. The basic concept is that of the construct validity of the test, the degree to which it measures some trait which really exists in some sense. Construct validity can be established only by convergence of several lines of evidence. Evidence for construct validity can be broken down into evidence that the test measures something systematic and evidence for the particular interpretation of what it measures. The degree of internal structure of the items and the magnitude of external correlations are the former, or psychometric, evidence; the nature of the structure, content of the items, and nature of the external relations are the latter, or psychological, evidence.

Test behavior is in the first instance responses to items. Such responses are both signs and samples. Because they are samples of behavior in general, they must be subject to the same laws as behavior. Because they are signs, inferences may be drawn from the organization of test responses to the organization of other behavior. Thus psychometrics must draw from but can also contribute to psychological theory.

There are three criteria for the construct validity of a test, mutually exclusive, exhaustive, and mandatory. These criteria are that the substance or content of the items shall be consistent with the proposed interpretation, that the structural relations of the items shall be consistent with the structural relations of non-test manifestations of the same trait, and that the external correlations of the test score shall not all be zero and shall be consistent with predictions

based on what is known of the postulated trait. Each item response is viewed as a bit of behavior rather than as a miniature measurement. Surplus meaning, psychological or psychometric, is avoided in interpreting item responses. In order for a test score to be useful, however, it must be demonstrated to have at once psychological and psychometric meaning, that is, the test score must be demonstrated to serve as a measure of some trait, subject, of course, to a margin of error.

The reasoning by which observations are parlayed into measurement may be summarized for the classical quantitative case. Several items which measure such a trait have greater probability of being answered in the positive direction, the greater the amount of the trait. A cluster of items measuring a common trait can be detected by the mutual intercorrelation of the items. The sum of several such items constitutes a score which tends to be greater, the greater the amount of the trait. The proportion of variance of the total score which is determined by the common trait can be large, even though for each item the proportion of variance determined by that trait is small. It is not assumed that the underlying trait or the observed scores are distributed according to the normal or indeed any other particular curve. That many observed distributions conform more or less to the normal curve may be a consequence of the fact that the scores are obtained by summing many items, each largely composed of error variance.

Full utilization of objective tests in clinical and in research settings depends on further development of a theory of test behavior and integration of that theory with the rest of psychology: what levels of personality, what impulses, defenses, and adaptive traits, are accessible to objective tests with different formal and content characteristics?

The foregoing considerations relate to theory of the first test. In order to interpret re-test or parallel form scores, there must also be available a theory of the second test, which requires study of secular trends in test behavior.[5]

B. METHOD

The program of test construction implied by the foregoing concepts may be summarized as follows:

1. The pool of items should be constituted so as to sample some area of content defined more broadly than the anticipated test. When feasible, all possible alternative constructs should determine the definition of the area of content. A wide and systematic sampling of the area of content, guided by the life-importance of subareas, is highly desirable.
2. The choice of the structural model for test construction should be based on what is known of the manifestations of the trait or type of trait measured in and outside of the test situation. Alternative structural models may be used with final selection based on empirical or theoretical criteria or a combination. The pool of items is administered to a normative sample and the best items are selected from the pool in conformity with the structural model chosen. Those items constitute the test or the scoring key. Degree of structure is ascertained by administering the test to a new sample.

[5] Dr. R. M. W. Travers has called my attention to his recent work showing the existence of secular trends within a single test administration for certain kinds of test materials and has raised the question of whether such tests are excluded from the discussion. The discovery of secular trends within a single test is not, of course, entirely new; similar observations have led to the virtual abandonment of the split-half method of computing test reliability. The external correlations of a test score are what they are regardless of trends within the several items which contribute to the score. There seems no reason, therefore, to exclude such tests from the discussion. A complete and rigorous treatment of score-structural theory, which has not been attempted in this monograph, certainly requires consideration of the problem Travers has raised, but the major points of the present discussion will not be affected.

3. The test score is then correlated with external (test and non-test) variables, both those that are expected to show relationship and those not expected to show relationship (sources of distortion).

The empirical findings of steps 2 and 3 are examined for concordance with alternative theories, or explanations, or constructs. The minimum criterion for acceptable validation is that the interpretation of the test scores be over-determined. If a separate principle or explanation must be invoked for each line of evidence, validation is not convincing.

The present view is a kind of operationism but certainly not the kind which states or implies that any set of operations defines a concept. Garner, Hake, and Eriksen [36] have suggested, in line with current versions of operationism in philosophy, that only *convergent* operations define a concept. The present paper is an attempt to make such a view explicit in relation to test construction.

C. APPLICATION

For a detailed example of the application of concepts similar to those developed in the present monograph, the reader is referred to a recent article by Jessor and Hammond [60]. They treat at length the use of the Taylor Anxiety Scale as an instrument of psychological theory.

A fragment of some research by Dr. Blanche Sweet and the writer will illustrate a number of points. A large number of items covering everyday problems of family life was collected and administered to a number of groups, particularly mothers of college students and girls of college age, many of whom were not, however, students.

Item 130 reads: "130a. After all the sacrifices parents make, teen-age children should be grateful to them. 130b. Teenagers cannot be expected to be grateful to their parents." *Ss* are required to choose one of the alternatives. A naive belief that behavior is directly translated into test response would lead one to expect that parents would tend to choose the former and adolescents the latter response. In fact, the trend is the other way, with the adolescent girls tending to choose the first alternative and the mothers tending to choose the second. The item belongs to a cluster whose predominant theme lies in the punitive versus permissive dimension. It is plausible that a group of young unmarried women would have a slightly more punitive and disciplinarian attitude than a matched group of mature women having had the experience of motherhood. A group of young women who had just enlisted in the Marine Corps were the most punitive group tested. A group of Vassar graduates, most of whom were also mothers, were the most permissive group tested.

The substance of the cluster makes sense, though many of the items obviously are not direct representations of observable behavior. A cumulative model seems adequate for the structure of this trait, that is, it seems reasonable to suppose that the wider the variety of situations in which the person chooses the more punitive alternative, the more punitive he is. Relationships with external variables, in this case group differences, are in concordance with the interpretation suggested by the content of the items.

Kinds of hypotheses that can be tested in such a course of test construction are:

(a) Those that concern organization of attitudes. Are individuals who are stern and punitive towards children equally rigorous in their demands on parents? Or, alternatively, are there "child-centered" and "parent-centered" people? Is authoritarianism a single trait? Or are there aspects of authoritarianism which vary more or less independently?

(b) Those that concern the relation of verbally expressed attitudes to personality structure. Granted that every single response reflects a variety of traits; do the clusters or themes among the items reflect consciously or unconsciously held attitudes? How superficial or fundamental in the

attitudes?

(c) Those that co
trait formation. For
that young inexperi
be more disciplinar
women is confirmed
be suggested. Do i
trol of their own
established tend to
capacity for reasor
hence believe in mo
of discipline?

Evidence for the
eses does not fit nea
The hypotheses ove
for them. None o
hypotheses has dir
verbal behavior. Th
conceived to lie ne
obvious or literal c
observable behavic
tudes nor in the c
of particular behav
test. The common e
metrics and naive
mental-theoretical
assume that only be
ing. Circumstances
ior largely unpredic
its propensities. Th
metrics, as of our
main within the be

The example il
closely related to
bach and Meehl's
of test validation
with the use of test
butions to psycholc
sented as a method
a method of testin
logical hypotheses.
this method of thec
with experimental

SUMMARY

Logically, the ki
achieves is indepe

defensive and adaptive structure are such attitudes?

(c) Those that concern the dynamics of trait formation. For example, if the finding that young inexperienced women tend to be more disciplinarian than more mature women is confirmed, other hypotheses will be suggested. Do individuals whose control of their own impulses is not firmly established tend to mistrust other people's capacity for reasonable impulse control, hence believe in more or less rigid patterns of discipline?

Evidence for the various kinds of hypotheses does not fit neatly into compartments. The hypotheses overlap, as does evidence for them. None of the three kinds of hypotheses has direct reference to nonverbal behavior. The validity of the test is conceived to lie neither in the degree of obvious or literal correspondence between observable behavior and expressed attitudes nor in the degree of predictability of particular behaviors on the basis of the test. The common error of classical psychometrics and naively operational experimental-theoretical psychology has been to assume that only behavior is worth predicting. Circumstances contrive to keep behavior largely unpredictable, however constant its propensities. The focus of our psychometrics, as of our psychology, should remain within the behaving person.

The example illustrates also a point closely related to a major point of Cronbach and Meehl's paper, that the process of test validation is virtually coterminous with the use of tests for substantive contributions to psychology. What has been presented as a method of test validation is also a method of testing some kinds of psychological hypotheses. Other hands can assay this method of theory testing in comparison with experimental alternatives.

SUMMARY

Logically, the kind of validity a test achieves is independent of the method of test construction. If one asks, however, how best to construct a test, the answer is that each kind of validity contains by implication a program of test construction. The programs implied by content validity and classical (predictive and concurrent) validity are not appropriate for major test construction projects. Only construct validity, which aims at measuring real traits, promises tests which will both draw from and contribute to psychology.

The lines of evidence which together establish the construct validity of a test refer to its content, its internal structure, and relation to outside variables. A single explanation or theory must encompass all evidence, for construct validation to be approximated.

Systematic factors affecting retests with the same or parallel form result in secular trends in test behavior. In general, secular trends must be assumed to exist; classical reliability theory assumes that they do not. A method of test construction based on construct validation can dispense with test-retest and parallel form reliability.

A psychometrics with construct validity as its central concept can be used as framework for viewing many recent contributions to psychometrics, some of which contribute to a construct-oriented psychometrics and others of which contrast with it.

REFERENCES

1. Adkins, D. C. The effects of practice on intelligence test scores. *J. educ. Psychol.*, 1937, 28, 222–231.
2. Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. *The authoritarian personality.* New York: Harper, 1950.
3. Anderson, T. W. On estimation of parameters in latent structure analysis. *Psychometrika*, 1954, 19, 1–10.
4. Bayley, N. On the growth of intelligence. *Amer. Psychologist*, 1955, 10, 805–818.
5. Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 1946, 11, 197–214.

tive design of psychological experiments. Berkeley: Univer. of Calif. Press, 1947.

8. Brunswik, E. Representative design and probabilistic theory in a functional psychology. *Psychol. Rev.*, 1955, 62, 193–217.
9. Campbell, D. T. The indirect assessment of social attitudes. *Psychol. Bull.*, 1950, 47, 15–38.
10. Carroll, J. B. Criteria for the evaluation of achievement tests. In *Proceedings, 1950 Invitational Conference on Testing Problems.* Princeton: Educ. Testing Service, 1951. Pp. 95–99.
11. Cattell, R. B. *Description and measurement of personality.* Yonkers-on-Hudson: World Book, 1946.
12. Chiang, C. L. On the design of mass medical surveys. *Human Biol.*, 1951, 23, 242–271.
13. Coombs, C. H. *A theory of psychological scaling.* Ann Arbor: Engineering Res. Inst., Univer. of Michigan, 1952.
14. Coombs, C. H. Theory and methods of social measurement. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences.* New York: Dryden, 1953. Pp. 471–535.
15. Coombs, C. H., & Kao, R. C. *Nonmetric factor analysis.* Ann Arbor: Engineering Res. Inst., Univer. of Michigan, 1955.
16. Cronbach, L. J. Statistical methods applied to Rorschach scores: a review. *Psychol. Bull.*, 1949, 46, 393–429.
17. Cronbach, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt*, 1950, 10, 3–31.
18. Cronbach, L. J. Report on a psychometric mission to Clinicia. *Psychometrika*, 1954, 19, 263–270.
19. Cronbach, L. J., & Gleser, G. C. Assessing similarity between profiles. *Psychol. Bull.*, 1953, 50, 456–473.
20. Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, 52, 281–302.
21. Cronbach, L. J., & Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika*, 1952, 17, 127–147.
22. Dorris, R. J., Levinson, D. J., & Hanf-
23. Dubois, P. H., Loevinger, J., & Gleser, G. C. *The construction of homogeneous keys for a biographical inventory.* Res. Bull. 52–18, Air Training Command, Human Resources Research Center, 1952, Lackland Air Force Base.
24. duMas, F. M. *Manifest structure analysis.* Missoula: Montana State Univer. Press, 1955.
25. duMas, F. M., Frost, C. H., & Rashleigh, C. H. A manifest structure analysis of information files. *J. clin. Psychol.*, 1956, 12, 139–143.
26. Edwards, A., & Coombs, C. H. A theory of psychological scaling. *Psychometrika*, 1954, 19, 89–91. (Review)
27. Elias, G. Self-evaluative questionnaires as projective measures of personality. *J. consult. Psychol.*, 1951, 15, 496–500.
28. Eysenck, H. J. Criterion analysis—an application of the hypothetico-deductive method to factor analysis. *Psychol. Rev.*, 1950, 57, 38–53.
29. Eysenck, H. J. *The scientific study of personality.* London: Routledge & K. Paul, 1952.
30. Fiske, D. W., & Rice, L. Intra-individual response variability. *Psychol. Bull.*, 1955, 52, 217–250.
31. Forer, B. R., & Tolman, R. S. Some characteristics of clinical judgment. *J. consult. Psychol.*, 1952, 16, 347–352.
32. French, J. W. *The description of personality measurements in terms of rotated factors.* Princeton: Educ. Testing Service, 1953.
33. Frenkel-Brunswik, E. Psychoanalysis and the unity of science. *Proc., Amer. Acad. Arts Sci.*, 1954, 80, 271–350.
34. Furst, E. J., & Fricke, B. G. Development and applications of structured tests of personality. *Rev. educ. Res.*, 1956, 26, 26–55.
35. Gaier, E. L., & Lee, M. C. Pattern analysis: the configural approach to predictive measurement. *Psychol. Bull.*, 1953, 50, 140–148.
36. Garner, W. R., Hake, H. W., & Eriksen, C. W. Operationism and the concept of perception. *Psychol. Rev.*, 1956, 63, 149–159.

chologists a *soc. Psychol*
39. Guertin, W A. I. Researc Intelligence *Bull.*, 1956,
40. Guilford, . tion psychol 3, 3–11.
41. Guilford, . *Psychol. Bu*
42. Guilford, Christense general rea 1956, 63, 1
43. Gulliksen, New York:
44. Guttman, tive data. 139–150.
45. Guttman, retest relial 255–282.
46. Guttman, scale and i *Measmt*, 1
47. Guttman, opinion m *et al., Mea* ton: Princ 46–59.
48. Guttman, ture of *metrika*, 1
49. Guttman for scale *Sociologic* Brunswick Pp. 410–
50. Guttman analysis: (Ed.), *Ma sciences.* Pp. 258–
51. Guttman factor an 173–192.
52. Haggard *the analy* den, 195
53. Harrow

37. GOODENOUGH, F. L. *Mental testing*. New York: Rinehart, 1949.
38. GRAYSON, H. M., & TOLMAN, R. S. A semantic study of concepts of clinical psychologists and psychiatrists. *J. abnorm. soc. Psychol.*, 1950, 45, 216–231.
39. GUERTIN, W. H., FRANK, G. H., & RABIN, A. I. Research with the Wechsler-Bellevue Intelligence Scale: 1950–1955. *Psychol. Bull.*, 1956, 53, 235–257.
40. GUILFORD, J. P. Some lessons from aviation psychology. *Amer. Psychologist*, 1948, 3, 3–11.
41. GUILFORD, J. P. The structure of intellect. *Psychol. Bull.*, 1956, 53, 267–293.
42. GUILFORD, J. P., KETTNER, N. W., & CHRISTENSEN, P. R. The nature of the general reasoning factor. *Psychol. Rev.*, 1956, 63, 169–172.
43. GULLIKSEN, H. *Theory of mental tests*. New York: Wiley, 1950.
44. GUTTMAN, L. A basis for scaling qualitative data. *Amer. sociol. Rev.*, 1944, 9, 139–150.
45. GUTTMAN, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, 10, 255–282.
46. GUTTMAN, L. The Cornell technique for scale and intensity analysis. *Educ. Psychol. Measmt*, 1947, 7, 247–280.
47. GUTTMAN, L. The problem of attitude and opinion measurement. In S. A. Stouffer, *et al., Measurement and prediction*. Princeton: Princeton Univer. Press, 1950. Pp. 46–59.
48. GUTTMAN, L. Image theory for the structure of quantitative variates. *Psychometrika*, 1953, 18, 277–296.
49. GUTTMAN, L. The Israel Alpha technique for scale analysis. In M. W. Riley, *et al., Sociological studies in scale analysis*. New Brunswick: Rutgers Univer. Press, 1954. Pp. 410–415.
50. GUTTMAN, L. A new approach to factor analysis: the radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. Glencoe, Ill.: Free Press, 1954. Pp. 258–348.
51. GUTTMAN, L. A generalized simplex for factor analysis. *Psychometrika*, 1955, 20, 173–192.
52. HAGGARD, E. A. *Intraclass correlation and the analysis of variance*. New York: Dryden, 1958.
53. HARROWER-ERICKSON, M. R. The value and limitations of the so-called "neurotic signs." *Rorschach Res. Exch.*, 1942, 6, 109–114.
54. HATHAWAY, S. R., & MCKINLEY, J. D. A multiphasic personality schedule: I. Construction of the schedule. *J. Psychol.*, 1940, 10, 249–254.
55. HAYS, D. G., & BORGATTA, E. F. An empirical comparison of restricted and general latent distance analysis. *Psychometrika*, 1954, 19, 271–279.
56. HORST, P. Pattern analysis and configural scoring. *J. clin. Psychol.*, 1954, 10, 3–11.
57. HOVLAND, C. I., & SHERIF, M. Judgmental phenomena and scales of attitude measurement: item displacement in Thurstone scales. *J. abnorm. soc. Psychol.*, 1952, 47, 822–832.
58. HUMPHREYS, L. G. Individual differences. In C. P. Stone & D. W. Taylor (Eds.), *Annual review of psychology*. Stanford: Annual Reviews, 1952. Pp. 131–150.
59. HUNT, H. F. Testing for psychological deficit. In D. Brower & L. E. Abt (Eds.), *Progress in clinical psychology*. Vol. 1, Sec. 1. New York: Grune and Stratton, 1952. Pp. 91–107.
60. JESSOR, R., & HAMMOND, K. R. Construct validity and the Taylor anxiety scale. *Psychol. Bull.*, 1957, 54, 161–170.
61. KELLEY, T. L. The future psychology of mental traits. *Psychometrika*, 1940, 5, 1–15.
62. KELLY, E. L., MILES, C. C., & TERMAN, L. M. Ability to influence one's score on a typical pencil-and-paper test of personality. *Charact. & Pers.*, 1936, 4, 206–215.
63. LAFITTE, J. Spearman's form of criterion analysis. *Brit. J. stat. Psychol.*, 1954, 7, 57–60.
64. LAZARSFELD, P. F. The interpretation and computation of some latent structures. In S. A. Stouffer, *et al., Measurement and prediction*. Princeton: Princeton Univer. Press, 1950. Pp. 413–472.
65. LAZARSFELD, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, *et al., Measurement and prediction*. Princeton: Princeton Univer. Press, 1950. Pp. 362–412.
66. LAZARSFELD, P. F. A conceptual introduction to latent structure analysis. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. Glencoe, Ill.: Free Press, 1954. Pp. 349–387.

67. LENNON, R. T. Assumptions underlying the use of content validity. *Educ. psychol. Measmt*, 1956, 16, 294–304.
68. LINDZEY, G. Thematic Apperception Test: interpretive assumptions and related empirical evidence. *Psychol. Bull.*, 1952, 49, 1–25.
69. LOEVINGER, J. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, 61, No. 4 (Whole No. 285).
70. LOEVINGER, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychol. Bull.*, 1948, 45, 507–529.
71. LOEVINGER, J. The attenuation paradox in test theory. *Psychol. Bull.*, 1954, 51, 493–504.
72. LOEVINGER, J. Effect of distortions of measurement on item selection. *Educ. psychol. Measmt*, 1954, 3, 441–448.
73. LOEVINGER, J. Some principles of personality measurement. *Educ. psychol. Measmt*, 1955, 15, 3–17.
74. LOEVINGER, J. The universe. *Amer. Psychologist*, 1955, 10, 399. (Abstract)
75. LOEVINGER, J., GLESER, G. C., & DUBOIS, P. H. Maximizing the discriminating power of a multiple-score test. *Psychometrika*, 1953, 18, 309–317.
76. LORD, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, 17, 181–194.
77. LORD, F. M. Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 1955, 20, 1–22.
78. LORD, F. M. Some perspectives on "The attenuation paradox in test theory." *Psychol. Bull.*, 1955, 52, 505–510.
79. LUBIN, A. A note on "criterion analysis." *Psychol., Rev.*, 1950, 57, 54–57.
80. LYKKEN, D. T. A method of actuarial pattern analysis. *Psychol. Bull.*, 1956, 53, 102–107.
81. MACCORQUODALE, K., & MEEHL, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.*, 1948, 55, 95–107.
82. MARSCHAK, J. Probability in the social sciences. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. Glencoe, Ill.: Free Press, 1954. Pp. 166–215.
83. MCQUITTY, L. L. A statistical method for studying personality integration. In O. H. Mowrer (Ed.), *Psychotherapy: theory and research*. New York: Ronald, 1953. Pp. 414–462.
84. MCQUITTY, L. L. Theories and methods in some objective assessments of psychological well-being. *Psychol. Monogr.*, 1954, 68, No. 14 (Whole No. 385).
85. MEEHL, P. E. The dynamics of "structured" personality tests. *J. clin. Psychol.*, 1945, 1, 296–303.
86. MEEHL, P. E. Configural scoring. *J. consult. Psychol.*, 1950, 14, 165–171.
87. MEEHL, P. E. *Clinical vs. statistical prediction*. Minneapolis: Univer. of Minnesota Press, 1954.
88. MEEHL, P. E. Wanted—a good cookbook. *Amer. Psychologist*, 1956, 11, 263–272.
89. MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.*, 1955, 52, 194–216.
90. MOWRER, O. H. "Q technique"—description, history, and critique. In O. H. Mowrer (Ed.), *Psychotherapy: theory and research*. New York: Ronald, 1953. Pp. 316–375.
91. OWENS, W. A. Item form and "false-positive" response on a neurotic inventory. *J. clin. Psychol.*, 1947, 3, 264–269.
92. PEAK, H. Problems of objective observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences*. New York: Dryden, 1953. Pp. 243–299.
93. RILEY, M. W., RILEY, J. W., JR., & TOBY, J. *Sociological studies in scale analysis*. New Brunswick: Rutgers Univer. Press, 1954.
94. ROGERS, C. R., & DYMOND, R. F. (Eds.) *Psychotherapy and personality change*. Chicago: Univer. of Chicago Press, 1954.
95. SCHAFER, R. *Psychoanalytic interpretation in Rorschach testing*. New York: Grune and Stratton, 1954.
96. SHERIF, M., & HOVLAND, C. I. Judgmental phenomena and scales of attitude measurement: placement of items with individual choice of number of categories. *J. abnorm. soc. Psychol.*, 1953, 48, 135–141.
97. STAR, S. A. The screening of psychoneurotics in the Army: technical development of tests. In S. A. Stouffer, *et al.*, *Measurement and prediction*. Princeton: Princeton Univer. Press, 1950. Pp. 486–548.
98. STEPHE techniq 498.
99. STEPHE *Q-techn* cago: U
100. STOUFF D. G., improvi *Quart.*,
101. TERMA *persona* 1936.
102. THORNI reliabili special eral nat 9, 189–
103. THURST *ysis*. C 1947.
104. THURST tions of 1948, 3
105. THURST in per *Measm*
106. THURST *measur* of Chic
107. TOBY, selectir tabulat *logical* Brunsv Pp. 33
108. TRAVE in the *Measn*
109. TUCKE velopi of voc chol.

98. STEPHENSON, W. Some observations on Q technique. *Psychol. Bull.*, 1952, 49, 483–498.
99. STEPHENSON, W. *The study of behavior: Q-technique and its methodology.* Chicago: Univer. of Chicago Press, 1953.
100. STOUFFER, S. A., BORGATTA, E. F., HAYS, D. G., & HENRY, A. F. A technique for improving cumulative scales. *Publ. Opin. Quart.*, 1952, 16, 273–291.
101. TERMAN, L. M., & MILES, C. C. *Sex and personality.* New York: McGraw-Hill, 1936.
102. THORNDIKE, E. L. Tests of intelligence; reliability, significance, susceptibility to special training and adaptation to the general nature of the task. *Sch. & Soc.*, 1919, 9, 189–195.
103. THURSTONE, L. L. *Multiple factor analysis.* Chicago: Univer. of Chicago Press, 1947.
104. THURSTONE, L. L. Psychological implications of factor analysis. *Amer. Psychologist*, 1948, 3, 402–408.
105. THURSTONE, L. L. The criterion problem in personality research. *Educ. psychol. Measmt*, 1955, 15, 353–361.
106. THURSTONE, L. L., & CHAVE, E. J. *The measurement of attitude.* Chicago: Univer. of Chicago Press, 1929.
107. TOBY, J., & TOBY, M. L. A method of selecting dichotomous items by cross-tabulation. In M. W. Riley, *et al.*, *Sociological studies in scale analysis.* New Brunswick: Rutgers Univer. Press, 1954. Pp. 339–355.
108. TRAVERS, R. M. W. Rational hypotheses in the construction of tests. *Educ. psychol. Measmt*, 1951, 11, 128–137.
109. TUCKER, L. R. Some experiments in developing a behaviorally determined scale of vocabulary. Paper read at Amer. Psychol. Ass., San Francisco, Sept., 1955.
110. TUCKER, L. R. The objective definition of simple structure in linear factor analysis. *Psychometrika*, 1955, 20, 209–225.
111. TUKEY, J. W. Discussion: symposium on statistics for the clinician. *J. clin. Psychol.*, 1950, 6, 61–74.
112. VERNON, P. E. *The structure of human abilities.* London: Methuen, 1950.
113. WECHSLER, D. *The measurement of adult intelligence.* (3rd Ed.) Baltimore: Williams and Wilkins, 1944.
114. WHERRY, R. J., & WINER, B. J. A method for factoring large numbers of items. *Psychometrika*, 1953, 18, 161–179.
115. WINDLE, C. Test-retest effect on personality questionnaires. *Educ. psychol. Measmt*, 1954, 14, 617–633.
116. WINTHROP, H. Semantic factors in the measurement of personality integration. *J. soc. Psychol.*, 1946, 24, 149–175.
117. WITTENBORN, J. R. Statistical tests of certain Rorschach assumptions: the internal consistency of scoring categories. *J. consult. Psychol.*, 1950, 14, 1–19.
118. WITTENBORN, J. R. Symptom patterns in a group of mental hospital patients. *J. consult. Psychol.*, 1951, 15, 290–302.
119. WITTENBORN, J. R., *et al.* A study of adoptive children. I. Interviews as a source of scores for children and their homes. *Psychol. Monogr.*, 1956, 70, No. 1. (Whole no. 408).
120. ZUBIN, J. The determination of response patterns in personality adjustment inventories. *J. educ. Psychol.*, 1937, 28, 401–413.
121. Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull. Suppl.*, 1954, 51, Part 2, 1–38.
122. *Webster's new collegiate dictionary.* (2nd Ed.) Springfield, Mass.: Merriam, 1951.