# THE CONCEPTS OF RELIABILITY AND HOMOGENEITY

C. H. COOMBS[1]
University of Michigan

## I. *Introduction*

THE literature of test theory is replete with articles on the computation and interpretation of indices of reliability. In them one finds surprisingly little common agreement or even mutual understanding (6). In more recent years the concept of homogeneity, with its indices, has been added, with the result that the confusion has increased. We shall make no effort in this paper to review and summarize this literature but shall attempt to do three things:

(1) point out what we regard as the fundamental sources of this confusion;
(2) provide a theoretical foundation on the basis of which this confusion might be resolved;
(3) point out the further steps that must be taken to develop the theory and practice of mental testing.

## II. *Sources of Present Confusion*

There are two fundamental sources[2] of confusion in present test theory: one is the assumptions by means of which we arrive at an interval scale (3), and the second is the identification of

[1] This paper is an extension to the area of mental testing of some of the ideas contained in a chapter in a general theory of psychological scaling developed in 1948–1949 under the auspices of the Rand Corporation and while in residence in the Department and the Laboratory of Social Relations, Harvard University. While the author carries the responsibility for the ideas contained herein, their development would not have been possible without the criticism and stimulation of Samuel A. Stouffer, C. Frederick Mosteller, Paul Lazarsfeld, and Benjamin W. White in a joint seminar during that year. Development of the theory before and after the sojourn at Harvard was made possible by the support of the Bureau of Psychological Services, Institute for Human Adjustment, Horace H. Rackham School of Graduate Studies, University of Michigan. A version of these ideas was presented in a 1949 APA symposium on Test Homogeneity and Test Validity.

[2] A Complete discussion of the fundamental difficulties in present test theory is to be found in Thomas (5).

our statistical indices with the concepts they are presumed to measure. These two basic difficulties are intimately related and are both associated with our attempt to model psychological measurement on physical measurement. Let us discuss them briefly, in turn.

Consider the manner in which data are obtained in the area of mental testing: The method used is the method of single stimuli, in which there is one response from each individual to each stimulus. These responses comprise our basic data, and consist of two piles of items for each individual. One pile has the items which the individual passed and the other pile those items which he failed. Note that there is no information in the data for a given individual pertaining to (1) how well he passed one item compared with another, or (2) how badly he failed one item compared with another, or (3) finally, how badly he failed one item compared with how well he passed another. The only way to obtain metric relations in data collected by the method of single stimuli is to put the information in the data by means of a priori statistical assumptions concerning, for example, the shape of the distribution function of the abilities of the individuals on the attribute in question. A normal distribution is usually what is assumed in test theory but even this is not applied in a thoroughgoing fashion.

To carry out the assumption fully (1) the percentage passing each item should be corrected for chance, then (2) converted to a sigma score, and (3) items at equal intervals on this sigma scale should be selected for a final form. This procedure is usually not rigorously adhered to because, in the first place, it makes little practical difference, in many instances, if the items are not precisely distributed in a discrete rectangular distribution on this sigma scale. But there is another reason why it is not insisted that this procedure should be rigorously adhered to, and that is because the assumptions which lead to a unit of measurement implicitly require the further assumption of perfect homogeneity. The distrust of the procedure is supported by the fact that the assumption of perfect homogeneity can usually, if not always, be shown to be violated, even in such crude data as that collected by the method of single stimuli. Unfortunately, to many this is simply regarded as one of the sources of error variance and not as a fundamental theoretical obstruction.

Thus, in the method of single stimuli as applied to mental testing we create an interval scale without any built-in or inherent test of its validity. Having such a scale, then, it is permissible to use certain properties of numbers, and we have available a variety of statistical procedures for the analysis of behavior. We must, of course, allow for error variance, much of which we have put there ourselves in assuming an interval scale, and, consequently, a statistical theory of error becomes necessary and plays a dominant role in test theory. This, then, is one major source of difficulty in the area of tests and measurements but, important as it is, it is not as fundamental as the second source. The difficulty arising from assumptions leading to an interval scale is of significance primarily to the empirical aspect of psychological testing rather than to the theoretical aspect.

The second source of difficulty, which we consider to be of prime theoretical significance, has, however, arisen from the use of an interval scale. Basically, this second source of confusion is the fact that we have had no fundamental *psychological* rationale underlying our concepts in test theory. Rather, we find an easy road to the concepts of test score, difficulty of an item, reliability and homogeneity via statistical definitions of indices dependent upon the existence of an interval scale. We set up these statistical indices based on operational procedures, then give names to them and act as if they have certain obvious psychological meanings. We have gained readily obtainable empirical indices but have paid for them in psychological ambiguity and imprecise meanings and interpretations. While relatively easy to compute and apparently readily susceptible to empirical study, an invalid assumption of an interval scale would vitiate even their numerical precision. Thus, we have not one but many indices of reliability, each determined in a different way, and hence each implying a different meaning. We do not have, independently, a quantitative definition of the concept of reliability, psychologically derived, with a unique interpretation. We have a variety of meanings for the concept of reliability, depending upon the index used. It is our thesis that the concept of reliability should have a unique psychological meaning quantitatively defined, and the various indices should then be regarded as different kinds of approximations to the concept. The challenge, then, would be to the experimenter to devise indices which are better measures of the concept.

## III. *A Psychological Rationale for the Concepts of Reliability and Homogeneity*

*The Fundamental Equation.*—We shall now attempt to sketch a theoretical psychological foundation for the derivation of quantitative definitions of certain concepts of test theory.

Consider the concept of the difficulty of an item. We all have intuitive notions as to what the psychological meaning of the difficulty of an item is. It means how hard it is for some one to pass it. But we identify the difficulty of an item with the percentage of people passing it. We thus have a number to represent the difficulty of an item which is the same number for all the people in the sample. Yet we know that for some people the item was so easy that they passed it, and for others it was so difficult that they failed it. It is apparent that we must have a definition of the difficulty of an item which will permit different values for different people. Of course, such a definition could still permit an *average* difficulty corresponding in principle to the conventional definition.

In order to develop a psychological rationale for the difficulty of an item let us consider an arithmetic problem. Let this arithmetic problem require that an individual know how to perform certain operations. The problem might involve addition and subtraction, the use of log tables, and a certain amount of reasoning. Its solution requires a collection of abilities, each to a certain degree and combined in a certain way. We may, for the sake of simplicity in discussion, lump this particular combination of abilities and call it a single ability. The problem then requires that every individual possess at least a certain amount of this ability in order to solve it. We shall call the quantity of an ability required for the solution of a problem the $Q$ value of that problem or that item.

Shall we regard this $Q$ value of an item as its difficulty? We might, if we wish, so define the difficulty of the item. But this is not psychologically satisfying, because if we ask individuals how difficult an item is, some will say that it is easy and some will say it is difficult. How can the item have one $\mathcal{Q}$ value and yet give rise to all this disagreement about its difficulty? Obviously it must be because these different individuals are making their judgments from different points of view. A mathematics major says it is easy; a grammar school student says it is hard. The point of view depends on the amount of this particular ability the person has. Of the particular ability demanded by the item, the amount possessed by an individual will be designated his $C$ value, representing his capacity.

We have now a hypothetical continuum on which is a $\mathcal{Q}$ value representing the amount of an ability required by the item from any individual to whom it is administered, and we have also a $C$ value on this same continuum for each individual who attempts the item. How, then, shall we represent the degree of difficulty that this item has for a particular individual? This might be done in a number of ways. We have chosen to use the ratio of $\mathcal{Q}$ to $C$ to represent the psychological value or difficulty of this item for that individual and have called this ratio $P$, and thus we have the simple equation:

$$\mathcal{Q} = PC \tag{1}$$

Obviously, the greater an individual's capacity the smaller proportion of that capacity is required or exercised in solving the problem and the easier it appears to him.

Each time ($h$) an individual ($i$) responds to a stimulus ($j$) here is a set of values which satisfy $\mathcal{Q}_{hij} = P_{hij}C_{hij}$. The most frequent objectives of psychological measurement are to determine something about the $\mathcal{Q}$ values of each member of a set of stimuli and the $C$ values of each member of a group of individuals.

But *note*, and this is significant to our later problem of metric, we do not observe $\mathcal{Q}$ values and $C$ values. Instead, what we observe are the $P$ values. Thus, if an individual passes an item, we know that on that particular ability the individual's capacity[3], $C_{ij}$, was greater than the quantity[3], $\mathcal{Q}_{ij}$, required to pass the item and hence the $P_{ij}$ value was less than one. In

[3] The subscript $h$ is *one* here.

the method of single stimuli, which is the method most used in mental testing, we can divide the items into two categories for each individual, those whose $P$ values were less than one for him, and those whose $P$ values were greater than one[4]. From such data on several individuals we want to extract what information they contain about $\mathcal{Q}$ and $C$ values. If we refuse to make the assumptions which lead to an interval scale, exhaustive analysis of these data would yield, at best[5], the *order* of the stimuli, (the $\mathcal{Q}$ values) and the *order* of the people (their $C$ values).

We might digress for a moment to point out that with other methods of collecting data, such as the method of rank order, the method of paired comparisons, and the method of triads, we are able to collect, successively, much more information about the $P$ values of stimuli for each individual and hence learn more about $\mathcal{Q}$ values and $C$ values than we do from the method of single stimuli used in mental testing. Curiously enough it appears that we are going to be able to go further, with fewer assumptions, in the area of so-called qualitative attributes than in the area of mental testing.

*The Variance of an Individual's Score.*—Imagine now that we have a stimulus or test item and a group of individuals who respond to it. Each individual's response to the item provides a $P$ value. Of course we do not know the exact magnitude of a $P$ value, we know only whether it is less than one or greater than one, that is, whether the individual passed or failed the item. But this is a limitation of this method of collecting data. Let us imagine that we had a method which would give us the exact $P$ values. There would be, then, a distribution of $P$ values for the stimulus. This distribution represents the distribution of difficulties which the item has for the individuals in the group.

Each individual has one of the $P$ values in this distribution. Let us imagine that we could again administer this item to this same group of individuals *independently*[6] *of its previous ad-*

[4] We have avoided the complication introduced by the true-false and multiple-choice type of item in which an individual may get an item right by pure chance. There is no need for this complication from the point of view of constructing a theory.

[5] The conditions necessary are that $\mathcal{Q}_{hij}$ be constant over $h$ and $i$ and the $C_{hij}$ be constant over $h$ and $j$. For purposes of future generalization these constitute an extreme of class 1 conditions (1).

[6] Experimental independence.

*ministration.* Then, once again, each individual would have a $P$ value for this item. Would the successive $P$ values of an individual for the one stimulus be identical, even if the successive administrations were independent? This is a question of whether or not $P_{hij}$ is constant over $h$ for a given $i$ and $j$ and can only be answered by experiment. It might well be that in the case of one attribute, say arithmetic, these successive $P$ values would be almost constant for any given individual, whereas in the case of another attribute, say the aesthetic merit of a painting, the $P$ values might be greatly variable. In this latter case we would expect the $P$ values to be variable if the individual was not too clear as to just what he meant by aesthetic merit and hence used different criteria in successive evaluations of the painting. Thus, if the continuum is intrinsically different at different times, both the $\mathcal{Q}$ values of the stimulus and the $C$ values of the individual would be variable for the same *nominal* trait, like aesthetic merit, because the exact composition of the trait was variable.

We have conceived, now, of each individual in a group having responded a number of times to a stimulus and, hence, for each individual, $i$, there is a distribution of $P_{hij}$ values for the stimulus $j$. Let us now do the same thing for more stimuli, and imagine that there is for every individual a small distribution of his $P$ values for each stimulus within the total distribution of all individuals' $P$ values for each stimulus. The notation used is as follows:

$h = 1, 2, \cdots t$, (the number of times an individual responds to a stimulus)

$i = 1, 2, \cdots N$, (the number of individuals)

$j = 1, 2, \cdots n$, (the number of stimuli)

$$P_{ij} = \frac{1}{t} \sum_h P_{hij}$$

$$P_i = \frac{1}{nt} \sum_j \sum_h P_{hij}$$

$$P_j = \frac{1}{Nt} \sum_i \sum_h P_{hij}$$

$$\overline{P} = \frac{1}{Nnt} \sum_i \sum_j \sum_h P_{hij}$$

We are now in a position to define the status score, $S_i$ (2), of an individual as follows:

$$S_i = \frac{1}{nt} \sum_j \sum_h (P_j - P_{hij}) \tag{2}$$

or

$$S_i = \overline{P} - P_i \tag{3}$$

To put the status score of an individual in words, it is defined as the average difficulty of all the items for all individuals minus the average difficulty of all the items for him alone. Thus, we have made the score of the individual dependent upon the composition of the group of individuals of which he is a member. On this scale the average individual has a score of zero, and the better the individual the higher his score, since the easier the items are for an individual the smaller the proportion of his capacity is required to pass them and the larger would be $S_i$. Individuals below average would have negative status scores.

Inasmuch as, in principle, an individual has a score, an $S_i$, on every item every time he takes it, let us consider the composition of the variance of all these "scores" that get averaged together for a total score. If we designate by $V_i$ the total variance of an individual, we have

$$V_i = \frac{1}{nt} \sum_j \sum_h (P_j - P_{hij})^2 - S_i^2 \tag{4}$$

By adding and subtracting $P_{ij}$ inside the parentheses, expanding and collecting terms, the expression for $V_i$ becomes:

$$V_i = \frac{1}{nt} \sum_j \sum_h (P_{ij} - P_{hij})^2 + \frac{1}{n} \sum_j (P_j - P_{ij})^2 - [\frac{1}{n} \sum_j (P_j - P_{ij})]^2 \tag{5}$$

Making the following definitions,

$$D_i^2 = \frac{1}{nt} \sum_j \sum_h (P_{ij} - P_{hij})^2 \tag{6}$$

$$T_i^2 = \frac{1}{n} \sum_j (P_j - P_{ij})^2 - [\frac{1}{n} \sum_j (P_j - P_{ij})]^2 \tag{7}$$

we have

$$V_i = D_i^2 + T_i^2 \tag{8}$$

and $V_i$ is seen to have two components. These two components, $D_i$ and $T_i$, are of psychological significance. The first component, $D_i$, we call the individual's dispersion score and it represents the variability within an individual in repeatedly responding (independently) to the same stimulus, summed over all the stimuli. $D_i$ reflects an individual's internal consistency in responding repeatedly to the same stimuli. The contribution that is made to this component by each stimulus is essentially the precision of the individual's score on each item, and when summed over the items is a measure of the *precision* of the individual's total score on the test.

The $T_i$ component describes the variability of the individual's mean position within the group as the group passes from stimulus to stimulus. We call this score the individual's trait score.

Thus, we now have two concepts to represent the hypothetical behavior of an individual in response to repeated independent presentations of a set of items. We have the concept of a dispersion score which represents the precision of an individual's final total score on the test. And we have the concept of trait score which represents the stability of an individual's position within the group in passing from item to item.

*Reliability and Homogeneity.*—We shall now identify $D_i$ and $T_i$ with the concepts of reliability and homogeneity, respectively. We have here precise definitions of concepts from a psychological rationale such that the concepts may be manipulated mathematically and are susceptible to rigorous logic.

We shall use the terms $D_i$, dispersion score, precision, and reliability interchangeably; and the terms $T_i$, trait score, and homogeneity interchangeably. First, it is apparent from the mathematical definition of the concept of precision that it is a characteristic of an individual's behavior on the items comprising the test, and does not necessarily have the same value for every individual who takes a particular test. To put this in the more common terms of test theory, the reliability of a test or, as we define it, the precision of an individual's test score, may be different for every individual who takes the test. It is an approximation of unknown degree to assign the same coefficient to all individuals. This approximation, perhaps, would be reasonably close in the case of some mental tests, but in others the individual differences in $D_i$ might be considerable.

The relation between reliability and homogeneity is an interesting one. In principle we could construct a test which would have high precision, or reliability, and such that the items would have zero intercorrelations, or, for that matter, any values from plus one to minus one. Thus, if a man's score on one item was the number of children he has and on another item his cephalic index, and on a third item the number of clubs and societies he belongs to, his total score would have very high reliability. It does not necessarily follow, however, that the score means anything—that it represents a point on a continuum which is a psychological trait continuum. Obviously, then, the fact that one has high precision for a test score has no bearing on whether or not one is measuring some kind of meaningful psychological entity. If one takes a number of things which are qualitatively different and adds up the scores on these different things for each individual, then the total scores will be a set of numbers which may have the property of precision but will have no common quality.

Let us turn now to the trait score which we identify with homogeneity. This denotes the stability of an individual's position within a group. Such a measure would not be an exclusive property of an individual, as in the case of precision, but is a property of the group as a whole on the test, and hence $\mathcal{T}_i$ should be averaged over the individuals.

The significance of this concept lies in its indicating the degree to which the final total scores of individuals have some common quality or represent a psychological entity for the group. The expression for the trait score, $\mathcal{T}_i$, averaged over individuals, is essentially equivalent to the notion of correlation between items, except that it is expressed in terms of variance rather than correlation or covariance.[7]

Thus, if we have a test consisting of a number of items, each

[7] Another way of looking at $D^2_i$ and $T^2_i$ is by analogy with error variance and true variance in conventional test theory. The analogy between $D^2_i$ and error variance is justified. But $T^2_i$ is a variance generated by lack of homogeneity among the items. Hence, in the sense used here, the "true variance" would represent the degree to which the items failed to constitute an organized and integrated common trait.

from a different primary mental ability, we would expect the position of the individual within the group from item to item to be variable. This is on the premise that there are intra-individual differences in ability. On the other hand, if the test were a set of arithmetic items then the position of the individual within the group as it passed from item to item would probably be relatively stable and there would be a high degree of homogeneity. These two tests might well have equally high reliability but quite different homogeneities.

In principle, the two components $D_i$ and $T_i$ are independent and it is not difficult to imagine a test with perfect precision for all individuals, or perfect reliability, and with a degree of homogeneity anywhere from zero to perfect. On the other hand, in a probability sense, it would perhaps be much more difficult to construct a test with perfect homogeneity but with low precision. Such a relation is implicit in the reasoning behind the attempt to increase the reliability of a test by means of an item analysis against an internal criterion.

*Indices.*—We have reached a point now where we must consider again the distinction between the defined meaning of a concept and the index which presumably is a measure of the concept. What we have tried to do is to provide meaningful definitions of the concepts of precision and homogeneity but we have *not* provided an *index* for either one of these concepts. An index is simply a method of analyzing data to get certain information. Hence, in order to compute a meaningful index, the data must contain this information. Consider, for example, what is required of the data so that they will contain information about the precision of an individual's score. We can see that to get a measure of precision, that is, to compute an individual's dispersion score, requires repeated independent responses from him to the same item. The method of single stimuli conventionally used in mental testing does not provide such observations. Thus, it appears that with conventional testing methods an index of the reliability of a test score is indeterminate and there is no valid formula for reliability. On the other hand, the $T_i$ component of an individual's total variance requires only one observation per individual per stimulus and, hence, data collected by the method of single stimuli do contain information pertaining to the concept of homogeneity. But samples of size *one* are poor estimates of the mean of a distribution. Nevertheless, they can be used to get an estimate of the variance between distributions which is, however, contaminated by the variance within the distributions. The two components, $D_i$ and $T_i$, of the total variance cannot be separated in data collected by the method of single stimuli. In other areas, a method for collecting data like the method of paired comparisons or the method of triads does provide information pertaining to both components and it is possible in principle to measure them both.

Essentially, what we have done is to give the quantitative definition of concepts based on a psychological rationale precedence over the statistical procedure of computing an index and then arguing about what the index means. We have chosen to have meaningful concepts and to recognize that our measures of them are inadequate and approximate rather than to take the measures as experimental facts and try to give them psychological meaning with consequent ambiguity and controversy.

What is it, then, that we do get from our indices of reliability or homogeneity? It is apparent that we can have no clear index of either the precision of a test score or the homogeneity of a test from conventional testing methods. Every index designed to represent one or the other actually represents a joint effect. The various indices merely differ in the nature of their approximation, then, to $V_i$, the left hand side of equation (8), summed over all individuals.

Inasmuch as this $V_i$ is also the variance of an individual's score just as one of its components, $D_i$, is, one might ask what the difference is between them. The difference is that $D_i$, the variability within an individual, is the degree of precision of a score on the *test*. $V_i$, the left hand side of the equation, is the precision of the individual's score on the *attribute*, the domain which the sample of items represents. Obviously, the homogeneity of the items in a test has nothing to do with the precision of a score on the test. But, obviously, this same score, when regarded as an estimate of the individual's score on the domain or attribute of which the items constitute a sample, is dependent upon the homogeneity of the domain. The greater the homogeneity of the domain, the more alike will be the scores of an individual on successive samples of items from that domain.

## IV. *Next Steps*

As we see some of the implications of this for the further development of test theory, there appear to be three general alternatives, the first of which has two sub-alternatives:

1. Continue with the method of single stimuli as a method of collecting data. Then we can do one of two things: (a) make the necessary assumptions to achieve an interval scale and hence have numbers to manipulate,[8] or (b) drop the assumptions which lead to an interval scale and substitute Lazarsfeld's latent structure analysis (4). The first sub-alternative above is to continue in the conventional manner. This will permit easily accomplished empirical studies in which we could rarely have firm confidence and unambiguous interpretation. The second sub-alternative requires going in an entirely new direction. Lazarsfeld's latent structure analysis is a non-metric theory for the scaling of data collected by the method of single stimuli. Obviously, his theory could be taken over bodily by test theorists, although from a practical point of view there are still computational hurdles. Such difficulties, however, are mere mechanical limitations and are not defects of the theory.

2. A second general alternative is to discover or to develop a new method for collecting data which would enable us to put the items in rank order for each individual as to how well he passed them and how badly he failed them. If we could collect such data we would then have data which, with very simple assumptions, contain information about metric relations between stimuli and individuals (1).

3. A third alternative is to discover or to develop a new method for collecting data which would be equivalent to the method of paired comparisons. This would require repeated independent responses to each stimulus. Such data would contain information on the metric relations between stimuli and individuals, and, in addition, information on the two compo-

[8] A better sub-alternative here is to experimentally validate the assumptions of an interval scale if this is possible.

nents of precision and homogeneity, making a precise distinction between them possible.

## V. *Summary*

We have tried to show that the assumptions required for an interval scale and the identification of indices with concepts are serious obstacles to the further development of test theory. We have then developed a rational basis for defining the difficulty of a test item for an individual and, from this basis, developed mathematical expressions for the concepts of reliability and homogeneity. It was then made apparent that the measurement of reliability and homogeneity from the analysis of data collected by the method of single stimuli is not possible, as such data do not contain the necessary information. Several alternative directions for the further development of test theory are pointed out.

## REFERENCES

1. Coombs, C. H. "Psychological Scaling Without a Unit of Measurement." *Psychological Review*, (in press).
2. Coombs, C. H. "Some Hypotheses for the Analysis of Qualitative Variables." *Psychological Review*, LV (1948), 167–74.
3. Stevens, S. S. "On the Theory of Scales of Measurement." *Science*, CIII (1946), 677–80.
4. Stouffer, S. A. *et al. Measurement and Prediction*. Princeton: Princeton University Press, 1949.
5. Thomas, L. G. "Mental Tests as Instruments of Science," *Psychological Monographs*, LIV (1942), No. 3.
6. Thorndike, R. L. "Logical Dilemmas in the Estimation of Reliability." *National Projects in Educational Measurement*. Series I. Reports of Committees and Conferences. XI (1947), 21–40.