PSYCHOMETRIKA—VOL. 12, NO. 1
MARCH, 1947

TEST "RELIABILITY": ITS MEANING AND DETERMINATION

LEE J. CRONBACH
UNIVERSITY OF CHICAGO

The concept of test reliability is examined in terms of general, group, and specific factors among the items, and the stability of scores in these factors from trial to trial. Four essentially different definitions of reliability are distinguished, which may be called the hypothetical self-correlation, the coefficient of equivalence, the coefficient of stability, and the coefficient of stability and equivalence. The possibility of estimating each of these coefficients is discussed. The coefficients are not interchangeable and have different values in corrections for attentuation, standard errors of measurement, and other practical applications.

The literature of testing contains many discussions of test reliability. Each year, new formulations are offered, and new procedures for estimating reliability are championed. There appears to have developed no universally accepted procedure, and several writers have attributed this difficulty to the diversity of definitions for reliability now in use. It has often been suggested that perhaps the only effective way to resolve the conflicts among contending viewpoints is to replace the term "reliability," recognizing that it covers not one, but several concepts. The present paper attempts to restate the conflicting concepts and assumptions now current, and to offer a scheme for separating the various aspects of dependability of measurement.

The physical scientist generally has expressed the accuracy of his observations in terms of the variation of repeated observations of the same event. The mean of the squared deviations of these observations about the obtained mean is the "error variance." This is a measure of precision or reliability. If for the present we regard reliability as the consistency of repeated measurements of the same event by the same process, two fundamental differences between the problem of the physical scientist and the psychologist appear. The physical scientist makes two assumptions, both of which are adequately true for him. First, he assumes that the entity being measured does not change during the measurement process. By controlling the relevant conditions—and he usually knows what these conditions are and can control them—he can hold nearly constant the length of a rod or the pressure of a gas. When measuring a variable quantity, where his assumption is no longer valid, he abandons the method of successive observations and employs instead simultaneous observations. The psychologist cannot obtain simultaneous measurements of behavior, yet the quantities that interest him are always variable and the method of successive measurements requires an impossible assumption. The psychologist may wish to measure a hypothetical constant (aptitude, or a limen), but all he can ever observe is behavior, which is always shifting. It is one thing to test the accuracy of measurement of a quantity, quite another to test whether that quantity is constant. Judgment on the second question must await judgment on the first.

The second assumption of the physical scientist is that his measurements are independent. If one rules out his remembering prior measurements, this assumption can usually be made true. Successive measurements of psychological quantities are rarely independent, however, because the act of measurement may change the quantity. London (**10**) has recently described this difficulty by the physicist's term *hysteresis*.

The reliability of a test score has generally been defined in terms of the variation of scores obtained by the individual on successive independent testings. Neither the assumption of constancy of true scores nor the assumption of experimental independence is realized in practice with most psychological variables; therefore, *the reliability of a test, as so defined, is a concept which cannot be directly observed.* If there is no standard of truth, it is fruitless to compare one estimate with another and debate which is more correct. But by various assumptions which usually cannot be tested, we obtain usable statistics which describe the test. *Different assumptions lead to different types of coefficients, which are not estimates of each other.* In particular, as many writers have noted, an estimate of the stability of a test score is not at all the same as an estimate of the accuracy of measurement of behavior at any one instant. Jenkins cites Franzen's comments on certain physiological measures which have high split-half "reliabilities" and low retest "reliabilities" (**6**). The measuring technique may be extremely accurate in reporting a biological instant in the life of an individual but not measure a stable characteristic of the individual.

Both the physicist and the psychologist encounter the problem of observer error. In sighting through a telescope or scoring an essay test, there is likely to be appreciable constant and variable error in observing. If one compares several judgments by the same observer, he includes the variable errors of observation with the errors of measurement. Hence, he studies the reliability of "this measuring instrument used by this man." If scores obtained by several observers in simultaneous measurements are pooled for comparison, the constant error of each man is included as a source of variation. This procedure studies the reliability of "this measuring instrument used by different men." Since the human takes part in the measurement, one cannot study the reliability of an instrument apart from the men who use it.

Types of "Reliability"

It is known that

$$r_{tt} = 1 - \frac{\sigma_e^2}{\sigma_t^2}, \tag{1}$$

where r_{tt} is the reliability coefficient, σ_e^2 is the hypothetical error variance—the mean of the squared deviations of all obtained scores for each person from the mean obtained score for that person—and σ_t^2 is the variance of the scores of all persons on all the hypothetical independent trials.

It is convenient to consider the possible definitions of error of measurement in terms of variance. Using a bi-factor pattern to describe a test,* the variance of scores from a single testing may be expressed as follows:

$$\sigma_1^2 = \sigma_g^2 + \sigma_{f_1}^2 + \sigma_{f_2}^2 + \cdots + \sigma_{s_1}^2 + \sigma_{s_2}^2 + \cdots + \sigma_{s_n}^2 + \sigma_{\varepsilon_1}^2. \tag{2}$$

The terms have the following meanings:

σ_1^2 is the variance of obtained scores;

σ_g^2 is the variance in the general factor (if any) represented in the test items;

$\sigma_{f_1}^2$, $\sigma_{f_2}^2$, etc., are the respective variances in the orthogonal group factors of undetermined number, each of which is represented in two or more items;

$\sigma_{s_1}^2$, $\sigma_{s_2}^2$, etc., are the respective "specificities" of the n items—the part of the reliable variance of scores on the items which cannot be assigned to common factors; and $\sigma_{\varepsilon_1}^2$ like the residual variance.

The referents for these factors may be illustrated in a hypothetical examination in psychology. The general factor might include general knowledge of psychology, reading ability, motivation, and

* Another factor pattern could be assumed without changing the basic argument (4, 7-9, 107).

other characteristics. Group factors might be related to knowledge of separate topics, mathematical skill required in only a few items, and so on. Each item taps, in addition, some specific knowledge not demanded by other items. The specificity variance accounts for individual differences in these elements. The remaining variance may include momentary inattention, guessing, and other random elements.

For reference, the formula will be rewritten thus:

$$\sigma_1{}^2 = \sigma_g{}^2 + \sum \sigma_f{}^2 + \sum \sigma_{s_i}{}^2 + \sigma_{\varepsilon_1}{}^2. \tag{3}$$

Consider now the scores obtained from a seкies of independent measurements of the same individuals using the same test.

$$\sigma_t{}^2 = \sigma^2_{\bar{g}_x} + \sum \sigma^2_{\bar{f}_x} + \sum \sigma^2_{\bar{s}_{ix}} + \sum \sigma^2_{g_x} + \sum \sum \sigma^2_{f_x} + \sum \sum \sigma^2_{s_{ix}} + \sigma_{\varepsilon_t}{}^2. \tag{4}$$

$\sigma_t{}^2$ is the variance of all obtained scores about the grand mean;

$\sigma^2_{\bar{g}_x}$ is the variance of the mean general factor scores of all individuals about the mean for all individuals—the between-persons variance in g;

$\sigma^2_{\bar{f}_x}$ is the between-persons variance in a group factor;

$\sigma^2_{\bar{s}_{ix}}$ is the between-persons variance in specificity on any item;

$\underline{\sum \sigma_{g_x}{}^2}$ is the sum over individuals of the variances of the general-factor scores for each individual about the mean for that individual—the within-persons variance;

$\sum \sum \sigma_{f_x}{}^2$ and $\sum \sum \sigma_{s_{ix}}{}^2$ represent the corresponding within-persons variances in the group factors and specificities, respectively; and

$\sigma_{\varepsilon_t}{}^2$ is the residual variance.

The between-persons variances represent, as in the case of the single trial, individual differences in the factors. The within-persons variances represent instability of scores for each individual, as a result of changes from test to test.

These formulations permit an exact statement of what a "reliability coefficient" represents. Apparently at least four fundamentally different meanings of reliability are current:

(1) The "error variance" may be permitted to include, in equation (4), the terms $\sum \sigma_{g_x}{}^2$, $\sum \sum \sigma_{f_x}{}^2$, $\sum\sum \sigma_{s_{ix}}{}^2$, and $\sigma_{\varepsilon_t}{}^2$. That is, instability is regarded as an error of measurement. This is the coefficient defined by the correlation from repeated independent administrations

of the same test. The assumption of constancy is made, since any change of score from trial to trial is treated as an error of measurement. If that assumption is true, the instability terms vanish, but such constancy in all the behaviors a test measures is highly unlikely.

(2) The "error variance" may be permitted to include, in equation (4), the terms $\sum \sigma^2_{\bar{s}_{ix}}$, $\sum \sigma^2_{g_x}$, $\sum \sum \sigma^2_{f_x}$, $\sum \sum \sigma^2_{s_{ix}}$, and $\sigma_{\varepsilon_t}^2$. Both instability and specificity are treated as errors. This is the "reliability" defined by the correlation between successive independent administrations of equivalent tests. Because different items are used in preparing equivalent forms, the specific-factor scores of individuals on the two tests will be uncorrelated. These, therefore, contribute to changes in score and are treated as error. If the tests do not represent the same group factors, at least part of $\sum \sigma^2_{\bar{f}_x}$ is also added to the error variance.

(3) The "error variance" may be permitted to include in equation (3), the terms $\sigma_{s_i}^2$ and $\sigma_{\varepsilon_1}^2$. This defines "reliability" as the correlation between two equivalent tests administered simultaneously. Instability is excluded from consideration, and no assumptions of constancy are made. Specific-factor variances are included in errors of measurement. Depending on the degree of equivalence, part of the group-factor variance may also be treated as error.

(4) The "error variance" may be restricted, in equation (3), to the term $\sigma_{\varepsilon_1}^2$. This is "reliability" defined as the self-correlation of a test (see below). No assumption of constancy is made, and independence is not involved. The specific factors remain the same from test to test and are added to the true-score variance. All real variables measured by the test are treated as quantities estimated, not as errors.

It may now be helpful to restate these definitions and to give them names for reference.

> *Definition* (1): *Reliability is the degree to which the test score indicates unchanging* individual differences in any traits. (Coefficient of stability).*
>
> *Definition* (2): *Reliability is the degree to which the test score indicates unchanging individual differences in the general and group factors defined by the test. (Coefficient of stability and equivalence).*

* This may be modified by requiring constancy over some specified period (one year, one day, etc.)

> *Definition* (3): *Reliability is the degree to which the test score indicates the status of the individual at the present instant in the general and group factors defined by the test.* (*Coefficient of equivalence*). Internal consistency tests are generally measures of equivalence. These coefficients predict the correlation of the test with a hypothetical equivalent test, as like the first test as the parts of the first test are like each other.
>
> *Definition* (4): *Reliability is the degree to which the test score indicates individual differences in any traits at the present moment.* (*Hypothetical self-correlation*).

These names are open to criticism, and better suggestions are in order. The important thing is to recognize that in the past all four of these and many approximations to them have been called "the reliability coefficient." No one of these is the "right" coefficient. They measure different things, and each is useful. What is important is to avoid confusing one with another, and using one as an estimate of another. It may be noted that reliability of a test can only be discussed in relation to a particular sample of persons.

The components of error variance under each definition imply that in practice some coefficients will be larger than others for a given test. If stability is not perfect, and if items contain some specificity loading, the hypothetical self-correlation will be greatest, and the coefficient of stability and equivalence will be the smallest of the four.

As Kelley states (7), the concept of reliability is meaningless unless one postulates that two measures of the same function exist. They may be successive measurements of a stable event, or simultaneous measurements of a unique event. But in regard to the non-repeating event which can be observed only once, reliability has only a theoretical interest. In fact, if one accepts a deterministic position, there is no "error" in a measurement of a unique event. The student's responses and his score are determined by many forces, and we do not know what they are; but the resultant of these forces is a particular act, and the act itself, at this instant and with these particular forces, is perfectly reliable. "Chance" and "error" are merely names we give to our ignorance of what determines an event.

All methods of studying reliability make a somewhat fallacious division of variables into "real variables" and "error." It is probbaly more correct to conceive a continuum between the instantaneous behavior which has an infinitesimal period, through states of longer duration, to the virtually constant individual differences. A test score is made up of all these "real" elements, each of which could be perfectly predicted if our knowledge were adequate. Reliability, according to this conception, becomes a measure of our ignorance of the real factors underlying brief fluctuations of behavior and atypical acts. Perhaps a new statistical method based on the non-Aristotelian conception of a continuum of realities will some day permit us to avoid the troublesome attempt to divide the continuum into "reality" and "error."

For the present, it appears to be necessary to retain the artificial separation. In thinking about the self-correlation of a test—the consistency with which it measures whatever it measures—we may class as chance effects all variables whose period of variation is shorter than the time required to take the test. Momentary fluctuations are therefore "errors," but shifts in fatigue, set, or skill having a longer cycle are possibly worth measuring.

Techniques of Estimation

Each method used in the past to study "reliability" may be associated with one of these definitions. The procedures requiring more than one trial will be discussed first.

Retest method. The retest method calls for giving the same test twice to the same group. The trials are supposed to be independent, but this may well not be true. Shift in relative scores is always treated in the error variance, not the true-score variance; the retest coefficient is therefore an estimate of the coefficient of stability. Failure to attain independent trials may make the estimate too high or too low.

Guttman (3, 263), in a complete reconsideration of reliability theory, defines reliability in terms of the stability of individual differences during a large number of "independent" retests. He shows that the reliability thus defined (a coefficient of stability) may be estimated by the correlation between two independent trials. His definition of independence will be discussed below.

Equivalent tests method. Two "equivalent" or "parallel" tests may be given, with any interval between, and their correlation determined. Experimental independence is assumed, despite the effect experience with one form may have on the second. Constancy is assumed, and all shifts in relative score are treated in the error variance. Specific-factor variances are treated in the error variance. This is therefore an estimate of the coefficient of stability and equivalence. Because the assumption of independence cannot be tested, it is never known whether the estimate is high or low. To interpret a coefficient involving equivalence, one must know how the tests are equivalent. If the tests are alike only in the general factor, group-factor variances are included as error, and the coefficient reflects the extent to which scores are determined by a stable general factor. Parallel tests should ordinarily have the same general and group factors. Were items in the two forms matched to test the same specific items of information or skill, the equivalent tests might to some degree include the same specific factors. The specific factors in the two tests could not be completely the same, however, unless the items were identical. The coefficient of equivalence is a property of a *pair* of tests and will vary according to the kind of similarity established in equating the tests. To the degree that parallel tests have the same general and group factors, the coefficient indicates the stability of performance in the general and group factors.

The split-half method. The widely used split-half method requires the correlation of half the items in the test with the remaining items. Cronbach has studied the effect of various splits upon the resulting coefficient (1) and has suggested the use of parallel splits, in which the two halves are made nearly equivalent (2). In the parallel split, each part represents the general factor and the group factors of the original test as well as possible. The half-tests should have equal standard deviations. The procedure makes no assumption of constancy, but does include the specific-factor variance as error variance. The split-half estimate is a coefficient of equivalence, estimating the correlation of simultaneously administered parallel tests, as like each other as are the halves of the test given. Any failure in splitting to obtain equivalent halves will tend to lower the correlation obtained. An assumption of experimental independence is made in considering the split-half correlation an estimate of the parallel-test correlation. In testing by parallel tests, the performance on one form is presumably independent of performance on the other. When items are presented together, however, there is always the possibility of spurious inter-item correlation due to item linkages and brief fluctuations of mood and attention.

Most random or odd-even splits do not represent all factors equally in both halves. If the assumption of experimental independence were valid, the correlation would therefore be an underestimate of the coefficient of equivalence. Guttman (3, 260) states that the corrected split-half coefficient is always a lower bound to "the reliability coefficient," no matter how the test is split. He cautions that this inequality is true only for an indefinitely large sample of persons. Sampling errors in practice preclude taking as one's coefficient the largest of many trial split coefficients. Guttman defines reliability in terms of repeated independent trials of the same (not equivalent) tests. By this definition, the split-half estimate, including specificity as an error of measurement, is a low one. The coefficient of equivalence is a conservative estimate of the hypothetical self-correlation.

The assumptions of the Spearman-Brown formula have been stated in various ways, and this has led to some confusion as to the applicability of the formula. The derivation hypothecates equivalent tests and predicts their correlation from the correlation of equivalent half-tests. Equivalence is the only assumption made, and in the derivation equivalence is defined by requiring equal standard deviations of the half-tests and by requiring that the hypothetical equivalent tests be just as similar as the half-tests ($r_{ab} = r_{aA} = r_{bB} = r_{AB}$). This defines equivalence so that all tests have the same common factor composition. It makes no direct assumption of the equivalence of pairs of items or of the unit-rank among the item intercorrelations.

The items of a test may be considered as a sample of some larger population. One may define the purpose of the test in terms of the population of items to be measured; the test fulfils this purpose insofar as the items are a *representative* sample of the population. Alternatively, one may consider the test as defined by its items, and think of the population as the entire group of items of which the sample is representative. The coefficient of equivalence (obtained by the parallel-test or internal consistency methods) correlates two samples of items and indicates the extent to which the variance in each may be attributed to common factors. The extent of common-factor loadings is the extent to which test scores are determined by "the population variable." If the samples to be compared must be representative, rather than random, it is necessary, in split-half procedures, to use the parallel split or a split according to a table of specifications.

The Kuder-Richardson formulas. A radical reformulation of the reliability problem was offered in 1937 by Kuder and Richardson (8). They proposed several alternative formulas which have been widely adopted. The original derivation has been criticized because of the numerous assumptions made, but other writers have developed the same formulas more directly. Perhaps the simplest derivation was published by Jackson and Ferguson (5, 74). They define reliability as a coefficient of equivalence, equivalence being defined by requiring that the two tests have equal variances and that the mean inter-item covariance within each test be equal and equal to the mean inter-item covariance between tests. If these assumptions are satisfied, the Kuder-Richardson formula (20) is an exact estimate of the coefficient of equivalence. This condition is a reasonable one when the items of a test are considered as drawn from a population of items all measuring a single general factor. If group factors are present, even though the two tests measures these group factors equally, then, $\overline{r_{i_j} S_i S_j} < \overline{r_{ij'} S_i S_{j'}}$,* and the Kuder-Richardson formula gives a conservative estimate of the coefficient of equivalence—how conservative one does not know.

The Guttman lower bounds. The latest statement of the problem is that published by Guttman in 1945 (3). He derives six formulas for estimating a coefficient from data obtained on a single testing, all the estimates being lower than the "true reliability" if the sample is sufficiently great. His estimate L_3 is identical to that from Kuder-Richardson formula (20), although the derivations are dissimilar. His L_4 is equivalent to the split-half coefficient. L_2, which uses item covariances, is an original formula more difficult to compute than L_3 and L_4. L_1, L_5, and L_6 are expected to have little practical importance.

Guttman defines error as the variation of the score of a person over a universe of independent trials with the same test. His crucial assumption, C_1 (3, 265-266), defines independence so that the score of a person on any item on any trial is experimentally independent of his scores on any other items. In practice, changes in motivation, function shift, and other variables cause items administered together to vary together. Guttman classes shifts in the variables measured as errors of measurement and therefore is estimating a coefficient of stability when he demonstrates that the correlation between two independent trials on a large population may be taken as equal to "the reliability coefficient" (3, 268).

In deriving lower-bounds formulas, Guttman deals with hypothetical independent retests in which the mean covariance of two items within trials equals the mean covariance of the same items between trials. Beyond this he makes no assumption. His definition of independence requires that there be no shift in the variables measured between trials; i.e., that the hypothetical trials be simultaneous. Since he is using identical tests simultaneously, he has defined reliability as *the hypothetical self-correlation.* His formulas lead to underestimates of that coefficient.

One may study the effect on Guttman's results if his assumption of independence within trials is denied. This may occur when one item influences the answer to another by giving a clue, by causing encouragement or discouragement, or by setting up a pattern among

* i.e., the mean inter-item covariance within tests is less than the mean inter-item covariance between tests.

the responses. In the derivation of L_1, the assumption leads to discarding a positive covariance term from the right member of (28). As a consequence, λ_1 and L_1 are greater than they would be without the assumption, and may overestimate the hypothetical self-correlation as defined. In the derivation of L_2, L_3, and L_4, the assumption is felt in (25), where a positive covariance term is dropped from the right member. Without the assumption,

$$\gamma_{x_g x_j} > \gamma_{X_g X_j}, \quad g \neq j,$$

and the inequality given in (37) may not hold. The remainder of the derivation therefore may lead to estimates higher than the hypothetical self-correlation, if the assumption of experimental independence of items does not hold.

This weakness is common to all estimates of reliability based on a single trial. Lindquist (9, 219) points out that in the split-half method the two halves are falsely assumed to be experimentally independent, and therefore he considers the split-half estimate spuriously high. [He, however, defines reliability as what we have called the coefficient of stability and equivalence (9, 216)]. In the Kuder-Richardson formula, as derived by Jackson and Ferguson, the same assumption of independence is made when the mean inter-item covariance between tests is taken as equal to the mean covariance within tests. If motivation, response sets, and other factors common to performance on the various items of a trial are considered part of the general or group factors measured by the test, their contribution to the inter-item correlation within a trial is rightly included in the estimate of accuracy of measurement. But momentary variations which cause random changes in item covariance should not be permitted to raise the estimate obtained. Any estimate of self-correlation or equivalence based on a single trial may be higher than the hypothetical self-correlation. It may be treated as a conservative or exact estimate only if we are willing to assume that the response to each item is an independent behavior, related to response on other items only because of significant conditions in the person tested.

Guttman makes the point that his split-half formula

$$L_4 = 2\left(1 - \frac{s_1^2 + s_2^2}{s_t^2}\right) \tag{5}$$

is superior to the Spearman-Brown formula in that it does not assume the two half-tests to have equal variance. His formula can be derived as an estimate of the coefficient of equivalence, according to the usual proof of the Spearman-Brown formula, except that equivalence is defined so that $\sigma_{a+b} = \sigma_{A+B}$, and $r_{aA}\sigma_a\sigma_A = r_{aB}\sigma_a\sigma_B = r_{Ab}\sigma_A\sigma_b = r_{bB}\sigma_b\sigma_B = r_{ab}\sigma_a\sigma_b$. This leads to a formula identical to Guttman's, or an equivalent form previously derived by Flanagan (see Kelley, 7) which is less readily computed. Values obtained using this formula are smaller (usually by a small amount) than the values from the Spearman-Brown formula, except where $s_a = s_b$. It appears that this formula should replace the Spearman-Brown procedure.

Summary

Four possible definitions of "reliability" have been considered. The hypothetical self-correlation requires independent simultaneous identical tests. For psychological variables this is a hypothetical situation, and no one has found an unbiased estimate of this coefficient. Guttman's formula L_2 would be a conservative estimate of the hypothetical self-correlation, save for the necessity of assuming that responses to one item are not influenced by responses to another item. Guttman's L_2 is ordinarily greater than the estimate from the Kuder-Richardson formula.

The coefficient of equivalence is lower than the hypothetical self-correlation. Kuder-Richardson formula (20) is an exact estimate of the coefficient of equivalence for tests where the item intercorrelation matrix has rank one; otherwise the estimate is conservative. This, however, like all estimates of equivalence, assumes experimental independence of items within one trial. The parallel-split method gives an estimate of the coefficient of equivalence. For an ideally large population, the highest split-coefficient is the best estimate, and estimates from other splits are conservative, save for the failure of independence of items.

The coefficient of stability is lower than the hypothetical self-correlation. It is estimated by the test-retest correlation, but carry-over from one test to another may cause the estimate to be faulty.

The parallel-tests correlation is an estimate of the coefficient of stability and equivalence. It may be unduly high if the two tests are not experimentally independent. Otherwise, the estimate will ordinarily be lower than the coefficient of stability or the coefficient of equivalence.

A simple table may indicate the different meanings of the various procedures. In Table 1, checks indicate the variances which are included in the error of measurement, according to each procedure. In the absence of sampling error, any estimate of reliability is less than the hypothetical self-correlation, assuming experimental independence. Every procedure assumes either the experimental independence of trials or of items within the trials. This condition is rarely satisfied, and any obtained coefficient may therefore be higher than the coefficient supposed to be obtained.

TABLE 1

Variances Included in Error Variance of a Test, According to Various Formulations of the Reliability Problem*

	General Factor Variance	Group Factor Variance	Specific Factor Variance	Instability Variance, General Factor	Instability Variances, Group Factors	Instability Variances, Specific Factors	Errors of Measurement of Items
Test-Retest				x	x	x	x
Parallel Test			x	x	x	x	x
Parallel Split			x				x
Random Split		x	x				x
Kuder-Richardson (20)		x	x				x
Guttman L_2							x†
Hypothetical Self-Correlation							x
Coefficient of Equivalence			x				x
Coefficient of Stability				x	x	x	x
Coefficient of Stability and Equivalence			x	x	x	x	x

* An *x* indicates that the variance indicated is included in the error of measurement by the procedure or definition listed at the left.

† In equations (31) and (43), Guttman sets up inequalities which overestimate the item error variance.

Practical Implications

No one "best" estimate of reliability exists. If one could validly make the assumption of stability between trials, and independence of trials, the test-retest correlation would be satisfactory. Frequently we must rely on single-trial estimates. Guttman's L_2 or a parallel-split used with his L_3 will in general give the highest coefficients. Where the test measures a single factor, the Kuder-Richardson formula (Guttman's L_4) should be as useful as the other two procedures.

In many situations, it is appropriate to seek a coefficient other than the hypothetical self-correlation. In correcting for attenuation, any of the coefficients described in this paper may be appropriate. Following the lead of Remmers and Whisler (**11**), one may distinguish between the "true instantaneous score" in a variable (related to the self-correlation or the coefficient of equivalence) and the "true score" in a trait (related to the coefficient of stability or of stability and equivalence). Sometimes one wishes to know the correlation between true scores in two traits postulated as stable over a period of time—"somatotype" vs. "temperament" is a typical problem. Here the appropriate coefficients for use in the attenuation formula are the coefficient of stability (if the trait is defined operationally by a specific test) or the coefficient of stability and equivalence (if the trait is defined by a family of similar tests). Other problems call for studying the relation between true instantaneous score in one variable (such as an aptitude test) and true score in another defined as stable (such as job performance). For this, the reliability of the former score would be based on a coefficient of equivalence (since the hypothetical self-correlation is not known), and the reliability of the latter would be based on one of the coefficients involving stability. The third possibility, and one of much theoretical importance, is a problem regarding true instantaneous scores in two variables, such as mood and performance. The correction for attenuation here requires use of two coefficients of equivalence.

Similar reasoning applies to the problem of estimating the significance of changes in test score. If the identical test is given both times, the coefficient of stability is appropriate. The hypothetical self-correlation, if known, would test whether a significant change in behavior had occurred, although this change might be due to normal diurnal fluctuation. The coefficient of stability tests whether the change is greater than that "normally" to be expected due to function fluctuation. If growth is measured by equivalent tests, a coefficient of equivalence, or of stability and equivalence, is relevant.

In evaluating a test, all four coefficients are of interest. For most purposes, one wishes to measure stable characteristics, so that a coefficient of stability is needed. For research purposes, however, a test having high instantaneous self-correlation or equivalence and low stability may be very satisfactory.

The coefficient of stability is an abstraction; in reality, there is an indefinitely large number of such coefficients, corresponding to various time intervals between tests. For meaningful use of such a coefficient, it must be defined as "the coefficient of stability over one week," or the like. The coefficient also depends on the conditions affecting the subject between testings. Strictly speaking, a coefficient of stability may be carried over to a new situation only when the time interval and the conditions between testings are similar to those under which the coefficient was obtained. The coefficient of stability would be better understood if research were available showing how the coefficient varies with increasing time lapse.

The following recommendations result from the analysis made above.

1. Reliability for psychological measurement can never be observed as in the physical sciences, where variables are practically constant and non-hysteretic. All estimates of reliability require as-

sumptions unlikely to be fulfilled.

2. Several coefficients numerically less than the hypothetical self-correlation can be estimated. A distinction between these various coefficients should be made; the writer proposes the names coefficient of equivalence, coefficient of stability, and coefficient of stability and equivalence.

3. The coefficient of equivalence may be estimated by the parallel-split method, using formula (5), Guttman's L_4. The Kuder-Richardson formula (20) underestimates this coefficient unless the test item matrix has rank one. Guttman's L_2 gives an underestimate of the hypothetical self-correlation which may or may not be higher than the coefficient of equivalence. All estimates of reliability or equivalence based on a single trial assume that test items are experimentally independent. To the extent that this is untrue, estimates may be erroneously high.

4. The coefficient of stability may be estimated by the test-retest method, with an undetermined error due to failure of independence. The coefficient of stability and equivalence may be estimated by the correlation of parallel tests, with a similar error.

5. In describing a test, the author should provide separate estimates of the coefficient of equivalence and the coefficient of stability. The time interval used in obtaining the coefficient of stability should be reported. If there are multiple forms, the coefficient of stability for each should be given.

6. In practice, the coefficient of equivalence or the coefficient of stability may be used meaningfully where the reliability coefficient is called for. The coefficients are not interchangeable and have different meanings in corrections for attenuation, standard errors of measurement, and like applications. The hypothetical self-correlation, showing the extent to which a test measures real but possibly momentary differences in performance, is more important to the theory of measurement than to the practical use of tests.

REFERENCES

1. Cronbach, L. J. A case study of the split-half reliability coefficient. *J. educ. Psychol.*, in press.
2. Cronbach, L. J. On estimates of test reliability. *J. educ. Psychol.*, 1943, **34**, 485-494.
3. Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, **10**, 255-282.
4. Holzinger, K. J. and Harman, H. Factorial analysis. Chicago: University of Chicago Press, 1941.
5. Jackson, R. W. B. and Ferguson, G. A. Studies on the reliability of tests. Toronto: Department of Educational Research, Bulletin No. 12, 1941.
6. Jenkins, J. G. Validity for what? *J. consulting psychol.*, 1946, **10**, 93-98.

7. Kelley, T. L. The reliability coefficient. *Psychometrika,* 1942, 7, 75-83.
8. Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika,* 1937, 2, 151-160.
9. Lindquist, E. F. A first course in statistics. Boston: Houghton-Mifflin, 1942.
10. London, I. D. Some consequences for history and psychology of Langmuir's concept of convergence and divergence of phenomena. *Psychol. Rev.* **1946, 53,** 170-188.
11. Remmers, H. H. and Whisler, L. Test reliability as a function of method of computation. *J. educ. Psychol.,* 1938, **29,** 81-92.