PSYCHOMETRIKA—VOL. 2, NO. 3
SEPTEMBER, 1937

THE THEORY OF THE ESTIMATION OF TEST RELIABILITY

G. F. KUDER AND M. W. RICHARDSON
The University of Chicago

The theoretically best estimate of the reliability coefficient is stated in terms of a precise definition of the equivalence of two forms of a test. Various approximations to this theoretical formula are derived, with reference to several degrees of completeness of information about the test and to special assumptions. The familiar Spearman-Brown Formula is shown to be a special case of the general formulation of the problem of reliability. Reliability coefficients computed in various ways are presented for comparative purposes.

The reliability coefficient is of interest because it gives, by the simple assumption that a test score has two components, viz., true score and variable error, an (indirect) estimate of the random error variance present in an obtained test score variance. No matter how computed, the reliability coefficient is only an *estimate* of the percentage of the total variance that may be described as true variance, i.e., not due to error.

The usual methods of estimating test reliability are too well known to justify description here. These methods differ in such a fashion that no close estimate can be made of the results of one method, knowing the estimate obtained by another method. It is always desirable, even necessary, for the investigator to state how he made his estimate of the reliability coefficient.* The retest coefficient on the same form gives, in general, estimates that are too high, because of material remembered on the second application of the test. This memory factor cannot be eliminated by increasing the length of time between the two applications, because of variable growth in the function tested within the population of individuals. These difficulties are so serious that the method is rarely used.

Although the authors have made no actual count, it seems safe to say that most test technicians use the split-half method of estimating reliability. This method involves an arbitrary division of the test

*The critical reader will reflect that, in addition, the investigator must report the range, or better, the variance of the group tested. The present study is not concerned with that matter.

into two parts, and the computation of the correlation-coefficient of the two sets of scores thus derived. The correlation coefficient thus obtained is taken as an estimate of the reliability of either half, and the Spearman-Brown formula for double length is then used to estimate the reliability coefficient of the whole test. The split-half method is commonly supposed to give estimates that are too high; this is an uncertain generalization unless one has some definitely defensible standard. A more pertinent observation about the split-half coefficient is that *it is not a unique value.* There are $\frac{n!}{2(\frac{n}{2}!)^2}$ different ways of dividing a test of n items into two halves. Each one of these ways of splitting the test gives its own estimate of the reliability coefficient.* True enough, not all these ways of splitting are equally defensible on *a priori* grounds. It remains true, however, that there are large fluctuations in the value of the reliability coefficient as obtained from difdiferent ways of constituting the two halves.†

The supposedly best method of estimating the reliability coefficient is to find the correlation between two *equivalent* forms, given at the same time. The crux of the matter here is *equivalence.* Actually the difficulties discussed in connection with the split-half coefficient still apply, in perhaps smaller degree. Again, there is no unique value of the reliability coefficient. In the quest for equivalence, the shift of items from one form to the other will affect the magnitude of the coefficient. In this situation, there are $\frac{(2n!)}{2(n!)^2}$ different coefficients, again not equally defensible.

In view of the limitations briefly described in the foregoing, the authors present certain deductions from test theory which lead to unique values of the reliability coefficient.‡ The least exact approximation we shall describe involves assumptions no more unreasonable

*With certain assumptions as to the distribution of inter-item correlations it would be possible to estimate, theoretically, the expected distribution of reliability coefficients thus to be computed. The most representative value (perhaps the mean) could then be taken as the best estimate and the problem thus solved. It is likely, however, that the solution would be enormously complicated by the possibility that the matrix of inter-item coefficients would have a rank greater than one. See Mosier, Charles I., "A Note on Item Analysis and the Criterion of Internal Consistency," *Psychometrika*, 1936, **1**, pp. 275-282.

†Brownell Wm. A., "On the Accuracy with which Reliability May Be Measured by Correlating Test Halves," *J. Exper. Educ.*, 1933, 1, pp. 204-215.

‡It should be mentioned that the main outlines of the simple argument in this article were derived independently by the two authors. In a chance conversation it developed that the two had reached similar conclusions by methods similar in principle.

than those basic to the Spearman-Brown formula. Any one of the formulas will give a unique estimate of the coefficient in all situations to which it is applicable. In certain cases, the commonly calculated parameters of the test score distribution will afford, in two minutes of time, a fairly good estimate of the reliability coefficient.

We shall consider a test variable t made up of n unit-weighted items applied to a population of N individuals. In the general case, we shall allow for the possibility of the inter-item coefficients varying between their possible limits, and also for varying proportions of correct answers; items need not be equally difficult or equally correlated with other items. This enables us to state the formally complete and theoretically most exact method of estimating the reliability of test t.

CASE I.

The data required are the number of items in the test, the difficulties of the items, the inter-item correlations, and the standard deviation of the total test. In one of the possible solutions suggested it is assumed that the matrix of inter-item correlations has a rank of one.

The correlation between two forms of a test is given by

$$r(a+b+\cdots+n)(A+B+\cdots+N) = \frac{r_{aA}\sigma_a\sigma_A + r_{aB}\sigma_a\sigma_B + \cdots + r_{n(N-1)}\sigma_n\sigma_{N-1} + r_{nN}\sigma_n\sigma_N}{[\sigma_a^2+\sigma_b^2+\cdots+\sigma_n^2+2(r_{ab}\sigma_a\sigma_b+r_{ac}\sigma_a\sigma_c+\cdots+r_{n(n-1)}\sigma_n\sigma_{n-1})]^{\frac{1}{2}}\times [\sigma_A^2+\sigma_B^2+\cdots\sigma_N^2+2(r_{AB}\sigma_A\sigma_B+r_{AC}\sigma_A\sigma_C+\cdots+r_{N(N-1)}\sigma_N\sigma_{N-1})]^{\frac{1}{2}}} \quad (1)$$

in which $a, b, \cdots n$ are items of the test, and $A, B, \cdots N$ are corresponding items in a second hypothetical test. Equivalence is now defined as interchangeability of items a and A, b and B, etc.; the members of each pair have the same difficulty and are correlated to the extent of their respective reliabilities. The inter-item correlations of one test are the same as those in the other. These relationships constitute the operational definition of *equivalence* which is to be used.*

By this definition of equivalence, the two expressions in the denominator of equation (1) are identical. It may then be seen that the numerator and denominator are the sums of the entries in square tables which are the same except for the entries in the principal diagonals. The entries in the principal diagonal of the numerator are the reliabilities of the items multiplied by their variance, while the entries in the diagonal of the denominator are merely the variances of the items. The formula for test reliability then becomes:

*It should be noted that this definition of equivalence is more rigid than the one usually stated.

$$r_{tt} = \frac{r_{aa}\,\sigma_a^2 + r_{bb}\,\sigma_b^2 + \cdots + r_{nn}\,\sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + r_{ac}\,\sigma_a\,\sigma_c + \cdots + r_{n(n-1)}\,\sigma_n\,\sigma_{n-1})}{\sigma_a^2 + \sigma_b^2 + \cdots + \sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + r_{ac}\,\sigma_a\,\sigma_c + \cdots + r_{n(n-1)}\,\sigma_n\,\sigma_{n-1})}. \tag{2}$$

The denominator of equation (2) is simply the expression for the variance of the sum of the items a to n, when each item is given a score of one. We can therefore substitute σ_t^2, the obtained variance of test scores, directly in the denominator, and also in the numerator by use of a suitable correction.

In order to write the numerator term, we must adjust the variance for the fact that the entries in the diagonals of the numerator and denominator tables are different. We therefore subtract from the obtained variance the sum of the variances of the items $(\sum_1^n pq)$ and substitute the sum of the products of the variance and reliability of each item $(\sum_1^n r_{ii}pq)$. The variance of any item i is p_iq_i.

The formula then becomes

$$r_{tt} = \frac{\sigma_t^2 - \sum_1^n pq + \sum_1^n r_{ii}pq}{\sigma_t^2}, \tag{3}$$

where σ_t^2 is the obtained test variance, $\sum_1^n pq$ is the sum of item variances, and $\sum_1^n r_{ii}pq$ is the sum of the products of item reliabilities and their variances.

Equation (3), while basic, is not adapted to calculations, because the r_{ii}'s are not operationally determinable except by use of certain assumptions. However, certain approximations are possible. If the inter-item correlations are available, two methods of estimating the n different values of r_{ii} suggest themselves. One is to use the average correlation of item i with the n—1 other items of the test as an estimate of the reliability of item i. This method, or other methods, of estimating the reliability of an item may be thought to be crude; however, it will be noted by reference to the square tables previously suggested that the r_{ii}'s comprise for a 100-item test only one per cent of the total number of entries whose values enter into the determination of the reliability coefficient of the whole test. Reasonable guesses as

to the values of r_{ii} would probaly not affect the final result very much, unless the tests were very short.

Another method is to estimate the unknown r_{ii} as the average computed from all the second-order minors of the matrix of inter-item correlations in which r_{ii} is the single unknown. By this method,

$$r_{ii} = \frac{\Sigma \frac{r_{ij}\, r_{ik}}{rjk}}{\frac{1}{2}(n-1)(n-2)}, \tag{4}$$

where i, j, and k are all different, and where the Σ means the sum of the separate determinations of r_{ii} from the $\frac{1}{2}(n-1)(n-2)$ minors. This method assumes that the matrix is of rank one, or that the test measures one function. This method would be justified only where n is fairly small.

CASE II.

The data required are the numbers of items in the test, the difficulties of the items, the item-test correlations, and the standard deviation of the test. It is assumed that the matrix of inter-item correlations has a rank of one.

A more usable approximation is adapted to those situations in which an item analysis giving values of item-test correlations has been made. If we care to assume that item and test measure the same thing (which, of course, we do when we put the item into the test), we may write

$$\frac{r_{it}}{\sqrt{r_{ii}\, r_{tt}}} = 1, \tag{5}$$

where r_{it} is the correlation between the item and the test, r_{ii} and r_{tt} are the reliabilities of item and test respectively.

Then

$$r_{ii} = \frac{r_{it}^2}{r_{tt}}. \tag{6}$$

Substituting $\frac{r_{it}^2}{r_{tt}}$ for r_{ii} in equation (3), we have

$$r_{tt} = \frac{\sigma_t^2 - \sum_1^n pq + \frac{\sum_1^n r_{it}^2 pq}{r_{tt}}}{\sigma_t^2}. \tag{7}$$

Solving for r_{tt}:

$$r_{tt} = \frac{\sigma_t^2 - \Sigma pq}{2\sigma_t^2} \pm \sqrt{\frac{\Sigma r_{it}^2 pq}{\sigma_t^2} + \left(\frac{\sigma_t^2 - \Sigma pq}{2\sigma_t^2}\right)^2}. \tag{8}$$

In practice, only the positive value of the radical in the right member of the equation is admissible. Equation (8) gives an estimate of the reliability coefficient in those situations in which the techniques of item analysis have been applied. In each case, Σ denotes summation over the n items.

CASE III.

The data required are the number of items in the test, the difficulties of the items, and the standard deviation of the test. It is assumed that the matrix of inter-item correlations has a rank of one and that all intercorrelations are equal.

In other situations, we may be willing to assume that the items are equally intercorrelated, but allow their difficulties to vary over a wide range. We shall proceed, therefore, to investigate this case. By assuming r_{ij} to be constant and equal to $\bar{r}_{ii}$ in equation (2) we have

$$r_{tt} = \frac{\bar{r}_{ii}\left(\sum_{i=1}^{n} \sqrt{p_i q_i}\right)^2}{\sigma_t^2}, \tag{9}$$

in which $\sqrt{p_i q_i}$ is the standard deviation of item i. Equation (9) gives an estimate of the reliability coefficient. An approximation to equation (9) is given by

$$r_{tt} = \frac{\bar{r}_{it}\Sigma\sqrt{pq}}{\sigma_t} \tag{10}$$

by assuming $\bar{r}_{ii} = \frac{\bar{r}_{it}^2}{r_{tt}}$, where $\bar{r}_{it}$ is the average item-test coefficient.

Since the test t is the sum of its items $a, b, \cdots n$, the variance of test scores is given by

$$\sigma_t^2 = \sigma_a^2 + \sigma_b^2 \cdots + \sigma_n^2 + 2(r_{ab}\sigma_a\sigma_b + r_{ac}\sigma_a\sigma_c + \cdots + r_{(n-1)n}\sigma_{n-1}\sigma_n, \tag{11}$$

in which $a, b, \cdots n$ are items of the test.

If all intercorrelations are assumed equal ($\bar{r}_{ii}$), and $\sqrt{p_i q_i}$ is used as the σ for an item,

$$\sigma_t^2 = (\Sigma\sqrt{pq})^2\bar{r}_{ii} - \Sigma pq\bar{r}_{ii} + \Sigma pq, \tag{12}$$

in which

$$\Sigma\sqrt{pq} = \text{sum of the } \sqrt{pq}\text{'s for items } a \text{ to } n,$$

and

$$\bar{r}_{ii} = \frac{\sigma_t^2 - \Sigma pq}{(\Sigma\sqrt{pq})^2 - \Sigma pq}. \tag{13}$$

Substituting for $\bar{r}_{ii}$ in formula (9)

$$r_{tt} = \frac{\sigma_t^2 - \Sigma pq}{(\Sigma\sqrt{pq})^2 - \Sigma pq} \cdot \frac{(\Sigma\sqrt{pq})^2}{\sigma_t^2}. \tag{14}$$

Again, all summations are over the items.

This formula is recommended for use when there is reason to believe that the inter-item correlations are approximately equal.

We shall digress slightly to illustrate the degree of approximation involved in the various steps. Let us suppose that, with reference to equation (9), we are in addition willing to assume equal standard deviations of items. With such an assumption, we have

$$r_{tt} = \frac{r_{ii} n^2 \overline{pq}}{\sigma_t^2}. \tag{15}$$

in which $\overline{pq}$ is the average item variance.

But

$$\sigma_t^2 = n\overline{pq}\,[1 + (n-1)\,r_{ii}], \text{ from (12)} \tag{16}$$

by similar assumptions. Substituting (16) in (15), we have

$$r_{tt} = \frac{n\bar{r}_{ii}}{1 + (n-1)\bar{r}_{ii}}. \tag{17}$$

Equation (17) is, of course, the familiar Spearman-Brown formula, which is predicated upon test length as the only variable affecting reliability, given a constant value of the reliability of the element.

It is now convenient to introduce another variant of equation (3), with assumptions similar to those involved in the Spearman-Brown formula.

From equation (12),

$$r_{ii} = \frac{\sigma_t^2 - n\overline{pq}}{(n-1)\,n\overline{pq}}, \tag{18}$$

since $\Sigma pq = n\overline{pq}$.

Substituting this value of r_{ii} in equation (15) we have

$$r_{tt} = \frac{\sigma_t^2 - n\overline{pq}}{(n-1)n\overline{pq}} \cdot \frac{n^2\overline{pq}}{\sigma_t^2}, \tag{19}$$

which simplifies to

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\overline{pq}}{\sigma_t^2}. \tag{20}$$

Equation (20) gives an estimate of the reliability of a test, knowing the number of items, the standard deviation, and the *average* variance of the items. This would not seem to be ordinarily a useful formula since it requires essentially the same basic data as formula (14), but involves one more approximation. Empirical evidence presented at the end of this paper, however, shows that reliabilities obtained by formula (20) do not for the tests used, vary more than .001 from those obtained from formula (14). Since formula (20) eliminates the necessity for computing $\sqrt{pq}$ for each item, it accomplishes a material saving in labor. It serves, too, as a basis for the formula recommended for use in Case IV.

CASE IV.

The data required are the number of items in the test and the standard deviation and mean of the total scores. It is assumed in this case that the matrix of inter-item correlations has a rank of one, that these correlations are equal, and that all items have the same difficulty.

Solution of formula (20) becomes greatly simplified if we make the rigid assumption that all items have the same difficulty. As the formula now stands it is necessary to obtain the average variance. The average variance ($\overline{pq}$) is equal to the product of average p and average q, ($\bar{p}\,\bar{q}$), if the items all have the same difficulty. In this case,

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\bar{p}\,\bar{q}}{\sigma_t^2}. \tag{21}$$

The average value of p may be easily obtained from the formula

$$\bar{p} = \frac{\Sigma X_t}{nN} = \frac{M_t}{n}, \tag{22}$$

when ΣX_t is the sum of the scores of N subjects on a test of n items, and M_t is the mean of the test scores.

The difference between equations (20) and (21) should be noted.

Equation (20) calls for the average of the item variances ($\overline{pq}$); equation (21) calls for the average of the item difficulties ($\bar{p}$) and this value subtracted from 1.00, ($\bar{q}$). When all items have the same difficulty, $\overline{pq}$ is equal to $\bar{p}\,\bar{q}$, but if there is variation in difficulty among the items, $\bar{p}\,\bar{q}$ becomes larger than $\overline{pq}$, and this discrepancy increases as the variation increases. This means that the estimate of reliability obtained by formula (21) is equal to or less than that obtained by formula (20). If Equation (22) is used to get an estimate of $\bar{p}$, the reliability coefficient can be quickly estimated from the mean, standard deviation, and the number of items. This formula may be regarded as a sort of foot-rule method of estimating test reliability without the necessity of splitting halves, rescoring twice, and calculating a correlation coefficient. According to theory and to the applications already made, the formula may be expected to give an underestimate of the reliability coefficient in situations not favorable for its application. If Equation (21) should give a higher value than the split-half, one would suspect the latter of being abnormally low because of some unfavorable way of splitting. The split-half Spearman-Brown coefficient cannot be regarded as the standard from which to judge other estimates. The split-half method involving use of the Spearman-Brown formula may produce estimates of reliability which are either too high or too low. Reliabilities obtained from the formulas presented here are never overestimates. When the assumptions are rigidly fulfilled, the figures obtained are the exact values of test reliability as herein defined; if the assumptions are not met, the figures obtained are underestimates.

It may be useful to suggest an interpretation of Equation (21) which has some bearing on the general problem of reliability. For r_{tt} to be positive, σ_t^2 must exceed $n\bar{p}\bar{q}$. Now $n\bar{p}\bar{q}$ is the variance of n equally difficult items when they are uncorrelated, by the familiar binomial theory.* Hence r_{tt} is positive for any average inter-item correlation that is positive. But negative reliability is inadmissible; hence only to the extent to which test items are positively intercorrelated will a test have reliability. It is implicit in all formulations of the reliability problem that *reliability is the characteristic of a test possessed by virtue of the positive intercorrelations of the items composing it.*

Table I presents a comparison of reliability coefficients computed

*Dunlap, J. W. and Kurtz, A. K., *Handbook of Statistical Nomographs, Tables and Formulas*, World Book Company, New York. Formula No. 46.

by equation (21) with a split half coefficient for various tests. The time of computation was approximately two minutes for each test, applying Equation (21).

TABLE I

Test No.	Nature	Range of values of p	$\bar{p}$	n	σ_t	Reliability Coefficient	
						By equation (21)	Split-half, Spearman-Brown
1	College Achievement	.05-.22	.156	50	6.56	.864	.880
2	"	.23-.40	.318	50	9.24	.891	.906
3	"	.41-.59	.522	50	10.96	.914	.923
4	"	.60-.77	.672	50	8.69	.872	.896
5	"	.78-.95	.852	50	6.57	.871	.888

Table II presents results from several formulas. As in Table I, three decimal figures are retained, merely to illustrate the differences obtained by the various formulas.

TABLE II

Test No.	Nature	Mean Score	n	σ_t	Reliability Coefficient, as estimated by			
					Case II Equation (8)	Case III Equation (14)	Case III Equation (20)	Case IV Equation (21)
6	multiple choice vocabulary	24.39	65	7.62	.823	.808	.808	.733
7	do	24.13	65	7.92	.839	.826	.825	.758
8	general information	.....	25	.....	.729	.716	.716	.714

The foregoing results are not intended to confirm the theory developed, but they may serve to illustrate the degree of divergences of results that may be expected in actual application. In comparing these estimates, it should be noted that all the tests are short; longer tests may be expected to give less variable estimates. Several algebraic variants are not here presented; they may be easily derived when their use is indicated. The choice of formula to be used in any actual situation will depend upon the amount of information about the components of the test, and upon the degree of accuracy desired. It is the belief of the authors that in many cases the quick estimate afforded by Formula (21) may be good enough for all practical purposes; if the items vary greatly in difficulty, Formula (20) appears to be adequate in any case.