PSYCHOMETRIKA—VOL. 18, NO. 1
MARCH, 1953

TEST RELIABILITY AND EFFECTIVE TEST LENGTH*

WILLIAM H. ANGOFF

EDUCATIONAL TESTING SERVICE

Measures of effective test length are developed for speeded and power tests, which are independent of the number of items in the test or of the time required for administration. These measures are used in determining reliability for (1) speeded and power tests, where a separately timed short parallel form is administered in addition to the full-length test; (2) power tests, where a subset of items is imbedded within the total test, parallel to the total test; and (3) power tests, where the subset of items is correlated with the complementary parallel subset in the test.

In a previous article, Cronbach (1) has pointed out that the characteristics of mental measurement that make the estimation of error particularly difficult are two-fold. First is that the very act of measuring produces a noticeable change in the object measured. The task of responding to test items, particularly items of a cognitive nature, is in itself a learning task, and on a second administration there is a variable positive bias in test performance which is generally attributed to increased test wisdom or to more specific acquaintance with test content. Second is the fact that uncontrolled changes during the process of measurement, as well as changes associated with growth and senescence (or learning and forgetting), also produce a changed performance on the second administration. In both cases the changed performance can be interpreted, in the context of test reliability, only as variable error unassociated with the reliability of the measuring instrument, and operating to reduce the size of the reliability coefficient.

In order to avoid attenuating the reliability coefficient with experimental error resulting from a second administration, methods have been developed for measuring reliability through the use of statistics taken from a *single* test administration. In general, two such methods have been made available, the Kuder-Richardson formulas and the split-half method (with Spearman-Brown correction for half length)—as well as variants of these later developed.

While these methods have yielded relatively satisfactory results for power tests, where sufficient time is given for all examinees to attempt all items, they have been considered totally inadequate for speeded tests.

*The writer gratefully acknowledges the assistance of Dr. Ledyard R Tucker in the formulation of some of the concepts presented in this paper. He wishes also to express his appreciation for the helpful comments of Dr. Harold Gulliksen and Dr. Frederic M. Lord in their review of the manuscript.

Guilford (3, 486) and Thorndike (7, 582), for example, have pointed out that an odd-even split of items in a purely speeded test would yield a correlation between test halves of unity, regardless of the reliability of the test. On the other hand, assuming that all examinees complete the first half of the test, a split of the first half against the second half would yield an indeterminate correlation, since the variability on the first half would be zero. In general, then, the computed reliability will be largely a function of the manner in which the test split has been made, and will tend not to reflect the actual reliability of the test in terms of the theoretical parallel-forms coefficient.

The Kuder-Richardson formulas are similarly inadequate for speeded tests. In speeded tests, where discrimination among examinees is made in terms of the differential number of items answered in a specified length of time, the inter-item covariances within a test are higher than they would be between parallel items on different forms of the test (7, 588). Since the reliability of the total test is a direct function of the reliabilities of the individual items (measured in this case in terms of inter-item correlations), the value of the reliability coefficient for the total test is thereby inflated.

In view of the inadequacies of the reliability formulas, it appears that there are at present no single-administration techniques for estimating the reliability of speeded tests. Guttman (5), in fact, maintains that reliability in general cannot be estimated from a single trial, and that all single-trial reliabilities are, in effect, lower bounds. Cronbach and Warrington (2) and Gulliksen (4) have developed lower-bound estimates of the reliability of speeded tests, but precise single-administration techniques are not available. Guilford (3, 486, 487) suggests the application of a split-half technique in which both test halves are given in separately timed administrations in immediate succession. (One of the difficulties of this method that first comes to mind is the matter of deciding on the appropriate time limits for the separate halves which would match the degree of speededness of the total test given in one administration.) The only alternative method is to devise an additional full-length parallel speeded test and to obtain an equivalent-form correlation. This procedure raises at least two problems: The first, discussed by Cronbach and Warrington, is the expense of constructing an alternate form solely for the purpose of providing reliability coefficients for a published test. The second problem relates to the questionable assumption that the parallel test is truly of the same effective length as the original test, merely because the numbers of items and the scheduled test times for the two tests are equal. In the case of speeded tests, variations in the amounts of time necessary to answer the items will cause substantial variations in the effective lengths of the tests.

The purpose of the present paper is first to suggest that the problem of economy in obtaining the reliability of a speeded test may be at least partly solved by administering a *short* parallel form in addition to the regular test.

Second, the purpose is to provide a measure of *functional* or *effective* test length and incorporate that measure into the reliability coefficient. In a later section of this paper, corresponding methods will be discussed for computing the reliability of unspeeded tests where the short parallel form is imbedded within the regular test, and only one administration of the test is given. In the latter case, the reliability is probably better interpreted as a lower-bound reliability or an index of internal consistency.

Case I. The determination of the reliability, r_{tt} , of test t from the correlation between test t and test i, a separately timed test, parallel to test t. While the more stringent case of speeded tests is treated, the method applies equally well to the case of unspeeded tests.

We shall consider that a short test, i, has been devised to parallel in function, level and spread of item difficulty, and items per unit of time a long test, t, which is speeded and for which a test reliability is to be determined. In connection with the requirement of parallelism it is assumed that the tests have been equated for spuriousness, in the sense that Cronbach and Warrington (**2**, 169) have used the term. In their paper they point out that in an unspeeded test "especial difficulty on one of the items neither increases nor decreases the person's probable standing on the remainder. But in a timed test, the person who gets stuck on one item may never reach the remainder of the items. It is this interdependence of items that introduces spuriousness." Finally, it is considered that, contained in test t, there are n tests, j, of effective length i, all parallel to test i. The correlation between tests i and t is given by:

$$r_{it} = r_{x_i(x_a+x_b+\ldots+x_j+x_k+\ldots+x_n)} = \frac{\sum_{j=1}^{n} C_{ij}}{\sigma_i\sigma_t} = \frac{n\overline{C_{ij}}}{\sigma_i\sigma_t}, \qquad (1)$$

where $\sum C_{ij}$ is the sum of the covariances $r_{ij}\sigma_i\sigma_j$ between test i and each of the parallel forms j of effective length i contained in test t. The value of n is the number of tests of effective length i contained in t, *or* the ratio of effective lengths, t to i.

The variance of test t may be written:

$$\sigma_t^2 = \sum_{j=1}^{n} \sigma_j^2 + \sum_{j=1}^{n}\sum_{k=1}^{n} C_{jk} = n\,\overline{\sigma_i^2} + n(n-1)\,\overline{C_{ik}}, \qquad (j \neq k). \qquad (2)$$

In general, throughout the development of the formulations to follow, it will be assumed that (a) average covariances involving parallel tests of equivalent length are equal, so that $\overline{C_{ij}} = \overline{C_{jk}}$; (b) any variance (or covariance) is equal to the average of all other variances (or covariances) involving parallel tests of equivalent length, so that $\sigma_i^2 = \overline{\sigma_i^2}$, $C_{ij} = \overline{C_{ij}}$, and that $C_{it} = \overline{C_{it}}$.

Then, solving (1) for $\overline{C_{ij}}$, substituting in (2) for $\overline{C_{ik}}$, replacing $\overline{\sigma_j^2}$ by σ_i^2, and solving (2) for n, we have

$$n = \frac{\sigma_t(\sigma_t + r_{it}\sigma_i)}{\sigma_i(\sigma_i + r_{it}\sigma_t)} . \tag{3}$$

Equation (3) yields a value of n which is determined not from the arbitrary ratio of the numbers of items in the two tests or from the ratio of time lengths, but from the data yielded by the test experiment itself. Particularly in speeded tests, neither the ratio of time lengths nor the ratio of numbers of items is suitable for estimating effective n. For one thing, assuming that no one completes the test, and that speed is the primary source of test variance, the distribution of test scores is highly sensitive to changes in total time limit as well as to changes in spuriousness (see above), but not at all sensitive to the addition of test items. Secondly, extraneous factors such as the period of warm-up at the beginning of the test would operate to reduce the effective test time in the short test to a greater extent, proportionally, than in the long test. Consequently, it would seem appropriate that a measure of effective test length be used in estimating reliability, such as that expressed in equation (3) rather than the ratio of the numbers of test items or the ratio of test times.

It may be of some interest to note that if $r_{it} = 1.00$, then $n = \sigma_t/\sigma_i$, and that if $r_{it} = .00$, then $n = \sigma_t^2/\sigma_i^2$. Consequently, we can establish that $\sigma_t^2/\sigma_i^2 \geq n \geq \sigma_t/\sigma_i$. It may also be observed from equation (3) that if the standard deviations of the tests are equal, then $n = 1$, and the tests are of equivalent length.

It will be convenient at this point to state the reliability of test t in terms of its correlation with test i. Consider that test t is correlated with a parallel test of equivalent length, composed of n tests of length i:

$$r_{tt} = r_{x_t(x_a+x_b+\ldots+x_i+\ldots+x_n)} = \frac{\sum_{i=1}^{n} C_{it}}{\sigma_t^2} = \frac{n\,\overline{C_{it}}}{\sigma_t^2} , \tag{4}$$

where $\sum C_{it} = \sum r_{it}\sigma_i\sigma_t$. In accordance with the assumption of equal average covariances stated above, $C_{it} = \overline{C_{it}}$. Thus,

$$r_{tt} = \frac{nC_{it}}{\sigma_t^2} = \frac{nr_{it}\sigma_i}{\sigma_t} . \tag{5}$$

Substituting (3) in (5), we find

$$r_{tt} = \frac{(\sigma_t + r_{it}\sigma_i)r_{it}}{\sigma_i + r_{it}\sigma_t} . \tag{6}$$

Equation (6) gives a method for determining the reliability of a test from its correlation with a parallel test, not necessarily of the same length. In examining the practicability of equation (6) it is observed that this is the formula to be used when estimating the reliability of a test from the correlation of any two parallel tests, even those *presumed* to be of the same effective length. If the standard deviations of the two tests are equal, and the tests are of equivalent length, then the reliability, r_{tt} , is identical to r_{it} , the correlation between the two tests. However, if the standard deviations are unequal, and the tests are incorrectly presumed to be of equivalent length, then the correlation between the two tests will be different from the reliability of either test. In effect, the value of n must be considered and incorporated into the determination of reliability, and it would be necessary to decide beforehand whether the reliability of test t is to be determined, or the reliability of test i. If the standard deviations of the two tests are different, then different results will be found.

Particular emphasis should be given to the basic assumption inherent in the present formulations: *tests* i *and* t *must be parallel tests.* If that assumption is violated in a choice of a non-parallel test i, then the reliability of test t may well be grossly underestimated.

Finally, it may be observed that if (1) is substituted in (5),

$$r_{tt} = \frac{n^2 \overline{C_{ij}}}{\sigma_t^2} . \tag{7}$$

If it is assumed that $\overline{C_{ij}} = r_{ii}\sigma_i^2$, then

$$n = \frac{\sigma_t \sqrt{r_{tt}}}{\sigma_i \sqrt{r_{ii}}} . \tag{8}$$

The value of n is seen to be the ratio of *the standard deviations of true scores* in the (mutually exclusive) long and short tests.

Case II. The determination of the reliability, r_{tt} , of an unspeeded test, t, from the correlation between test t and a subset of items, test j, included in test t.

If test t is not speeded and the principal source of test variance lies in the differential abilities of the examinees to respond correctly to test items, then a single test administration is capable of yielding an internal consistency reliability coefficient. Let us consider that there exists and can be chosen a subset of items, test j, contained in test t, that parallel the parent test in function and difficulty. Further, consider that there are n such parallel subtests contained in test t, all mutually exclusive. Then, making use of the same assumptions of equivalence as were made for equation (1) above,*

*Except for equating the characteristic of spuriousness. See statement of Cronbach and Warrington quoted above.

we can state the correlation between test j and its parent test, t, as follows:

$$r_{jt} = r_{x_j(x_a+\ldots+x_j+x_k+\ldots+x_n)}$$

$$= \frac{\sigma_j^2 + \sum_{i=1}^{n} C_{ik}}{\sigma_j \sigma_t} \qquad (j \neq k)$$

$$= \frac{\sigma_j^2 + (n-1)\,\overline{C_{jk}}}{\sigma_j \sigma_t} . \tag{9}$$

We have observed that

$$\sigma_t^2 = n\,\overline{\sigma_j^2} + n(n-1)\,\overline{C_{jk}} . \qquad (j \neq k) \tag{2}$$

If we now solve (9) for $\overline{C_{jk}}$, substitute in (2), and replace $\overline{\sigma_j^2}$ by its equivalent, σ_j^2,

$$n = \frac{\sigma_t}{r_{jt}\sigma_j} . \tag{10}$$

Now solving equation (2) for $\overline{C_{jk}}$ and substituting in (7) for its equivalent, $\overline{C_{ii}}$,

$$r_{tt} = \frac{n(\sigma_t^2 - n\,\overline{\sigma_j^2})}{(n-1)\sigma_t^2} , \tag{11}$$

which is exactly parallel to Kuder and Richardson's formula (20). Finally, if σ_t^2 is substituted for its equivalent, $\overline{\sigma_j^2}$, and the value found in (10) is substituted for n,

$$r_{tt} = \frac{\sigma_t r_{jt} - \sigma_j}{r_{jt}(\sigma_t - r_{jt}\sigma_j)} . \tag{12}$$

Equation (12) gives the reliability of an unspeeded test, t, obtained from the correlation between t and its parallel subtest j, and their standard deviations.

Case III. The determination of the reliability, r_{tt}, of an unspeeded test, t, from the correlation between its complementary parallel parts, h and j, and their standard deviations.

It will be observed that equation (12) may be written as follows:

$$r_{tt} = \frac{r_{j(t-j)}}{r_{jt} r_{t(t-j)}} . \tag{13}$$

If $h \equiv t - j$,

$$r_{tt} = \frac{r_{hj}}{r_{jt} r_{ht}} = \frac{r_{hj}}{r_{j(j+h)} r_{h(h+j)}} , \tag{14}$$

and

$$r_{tt} = \frac{r_{hj}\sigma_t^2}{(\sigma_h + r_{hj}\sigma_j)(\sigma_j + r_{hj}\sigma_h)} . \tag{15}$$

The value of σ_t^2 may be taken from the following expression,

$$\sigma_t^2 = \sigma_h^2 + \sigma_j^2 + 2r_{hj}\sigma_h\sigma_j ,$$

and substituted in (15) to yield

$$r_{tt} = \frac{r_{hj}(\sigma_h^2 + \sigma_j^2 + 2r_{hj}\sigma_h\sigma_j)}{(\sigma_h + r_{hj}\sigma_j)(\sigma_j + r_{hj}\sigma_h)} , \tag{16}$$

so that all values used are taken from the subtest scores.

It may be noted from (16) that if the test split has been made in such a way as to produce parallel tests of equal effective length—that is, when h and j are equivalent tests and $\sigma_h = \sigma_j$, then

$$r_{tt} = \frac{2r_{hj}}{1 + r_{hj}} ,$$

which is the familiar Spearman-Brown correction for half length.

To complete the analogy between *i-exclusive-of-t* and *j-contained-in-t*: It is clear that the counterpart, for the "contained" case, of equation (8) (where n is expressed as the ratio of the standard deviations of true scores, t to i), is directly analogous to the "exclusive" case. If it is assumed that $r_{jj}\sigma_j^2 = r_{ii}\sigma_i^2$, then equation (8) may be restated:

$$n = \frac{\sigma_t\sqrt{r_{tt}}}{\sigma_j\sqrt{r_{jj}}} . \tag{17}$$

In the case of power tests, test length has usually been measured in terms of the number of items. However, if the items near the beginning of the test are correctly answered by everyone in the group, or if the items near the end of the test are correctly answered by no one in the group, then the test is obviously not effectively of the length arbitrarily assumed. Some measure of test length should be used such as that implied in (17), which takes into account the number of items effectively discriminating among the members of the tested group.

It may be argued that if the short test is ideally chosen with respect to level and range of item difficulty, then the value of n will remain constant, irrespective of the performance of the particular group. However, since the ideal is not achieved in practice, it is necessary to determine the value of n in the particular instance. In effect, the direct determination of effective n allows a greater degree of laxity in the choice of items for the subtest, but *does* become a necessary adjunct to the determination of reliability. Particularly important is the fact that the choice of items for the subtest need not be restricted by any arbitrary prior decision regarding its length, since its length would be determined in conjunction with the determination of the reliability. With that restriction removed, greater freedom can be devoted

to making the subtest truly parallel in function and distribution of item difficulty.

It may be well to repeat that throughout these formulations it is assumed that the subtest of items, j, is parallel to the long test, t. If this assumption is not met in practice, then the reliability of test t will be underestimated.

It will be of some interest to examine the relationship among r_{tt} , r_{ii} , and r_{it} , where test i is exclusive of t, and to compare that relationship with that found among r_{tt} , r_{jj} , and r_{jt} , where j is contained in t. If we consider the "exclusive" case first, we note in equation (1) that

$$r_{it} = \frac{n\,\overline{C_{ij}}}{\sigma_i \sigma_t} \,. \tag{1}$$

If we assume that $\overline{C_{ij}} = r_{ii}\sigma_i^2$, and substitute in (1) the value of n found in (8), then

$$r_{it}^2 = r_{ii} r_{tt} \,. \tag{18}$$

Equation (18) has otherwise been obtained by stating the correlation between parallel forms of the same test, adjusted for attenuation due to unreliability and considering that the correlation between true scores on parallel tests is equal to unity.

Considering the "contained" case, we note in equation (10) that

$$n = \frac{\sigma_t}{r_{jt}\sigma_j} \,. \tag{10}$$

If the value of n found in equation (17) is substituted in (10), it is found that

$$r_{jt}^2 = \frac{r_{jj}}{r_{tt}} \,. \tag{19}$$

It is observed in comparing (18) with (19) that the relationship among r_{it} , r_{ii} , and r_{tt} in the "exclusive" case is quite different from the relationship among r_{jt} , r_{jj} , and r_{tt} in the "contained" case. When the short test is exclusive of t, then r_{it}^2 is equal to the product of the reliabilities of the short and long tests (equation 18); when the short test is contained in t, then r_{jt}^2 is equal to the ratio of the reliabilities of the short and long tests (equation 19).

It can be shown that equations (18) and (19) are not inconsistent, if account is taken of the spuriousness in (19). Since tests i and j are parallel and of equal length, assume that $r_{ii} = r_{jj}$. Solving (18) for r_{tt} and substituting in (19), and also substituting r_{ii} for r_{jj} ,

$$\frac{r_{it}^2}{r_{ii}} = \frac{r_{ii}}{r_{jt}^2} \,. \tag{20}$$

Now substituting in (20) the values of r_{it}^2 and r_{jt}^2 found respectively in (1) and (10), and assuming that $\overline{C_{ij}} = r_{ii}\sigma_i\sigma_j$, equation (20) results in an identity.

It may be of interest to examine further the relationship between r_{it} and r_{jt}. If tests i and j are parallel and of equivalent length, as has been assumed throughout this development, it would appear obvious that $r_{jt} \geq r_{it}$, because of the spuriousness in r_{jt}. The degree of this spuriousness can be shown in that $r_{jt} = \sqrt{r_{ij}r_{tt}} + k$, where

$$k = \frac{\sigma_j(1 - r_{ij})}{\sigma_t} = \frac{1 - r_{ij}}{\sqrt{n + n(n-1)r_{ij}}} .$$

Assuming that the reliabilities r_{ii} and r_{jj} are equal, then it is clear that $r_{jt} \geq r_{it}$, and that the relationship $r_{jt} = \sqrt{r_{ij}r_{tt}}$ cannot hold unless $r_{ij} = 1.00$.

Consider that the correlation between true scores on j and t (j included in t) is unity. Then

$$r_{j_\infty t_\infty} = 1.00 = r_{(x_j - e_j)(x_t - e_t)} = \frac{\sum x_j x_t - \sum x_j e_t - \sum x_t e_j + \sum e_j e_t}{N\sigma_j\sigma_t\sqrt{r_{jj}r_{tt}}} , \tag{21}$$

where, for example, j_∞ and e_j are the true and error components of x_j, such that $x_j = j_\infty + e_j$.

Examining each term separately, we find

$$\frac{\sum x_j x_t}{N} = r_{jt}\sigma_j\sigma_t ,$$

$$\left.\begin{aligned} \frac{\sum x_j e_t}{N} &= \frac{\sum x_j(e_a + e_b + \cdots + e_j + \cdots + e_n)}{N} = \frac{\sum x_j e_j}{N} = \sigma^2_{e_j}, \\ \frac{\sum x_t e_j}{N} &= \frac{\sum (x_a + x_b + \cdots + x_j + \cdots + x_n)e_j}{N} = \frac{\sum x_j e_j}{N} = \sigma^2_{e_j} , \\ \frac{\sum e_t e_j}{N} &= \frac{\sum (e_a + e_b + \cdots + e_j + \cdots + e_n)e_j}{N} = \frac{\sum e_j^2}{N} = \sigma^2_{e_j} . \end{aligned}\right\}$$

Other terms go to zero.

Then

$$1.00 = \frac{r_{jt}\sigma_j\sigma_t - \sigma^2_{e_j}}{\sigma_j\sigma_t\sqrt{r_{jj}r_{tt}}} , \tag{22}$$

and

$$r_{jt} = \sqrt{r_{jj}r_{tt}} + \frac{\sigma_j(1 - r_{jj})}{\sigma_t} ;$$

and finally,

$$r_{jt} = \sqrt{r_{jj}r_{tt}} + \frac{1 - r_{jj}}{\sqrt{n + n(n-1)r_{jj}}} . \tag{23}$$

It has been observed that equation (18) holds only when the short and long tests are mutually exclusive, and that (19) is applicable when the short test is contained in the long test. The use of (18) when (19) is justified may well lead to questionable results. For example, it appears that in their Case II development, Kuder and Richardson (6) made inappropriate use of equation (18), since they were dealing with the correlation between an item and the test in which it was contained.* If their equation (3) is restated in the present notation (and renumbered *a*),

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

and $r_{tt}\, r_{it}^2$ is substituted for r_{ii} instead of r_{it}^2/r_{tt}, which they used, then

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

Solving for r_{tt}, we find

$$r_{tt} = \frac{\sigma_t^2 - \sum_1^n pq}{\sigma_t^2 - \sum_1^n r_{it}^2pq}, \tag{c}$$

instead of the equation (8) presented in their article,

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

A second instance in the Kuder-Richardson article of the inappropriate use of equation (18) appears in their Case III development, in the step from their equation (9) to equation (10). They assumed that $r_{ii}\, r_{tt} = r_{it}^2$, which does not hold if item j is contained in test t. If their equation (9) is restated in the present notation (and renumbered *d*),

$$r_{tt} = \frac{\overline{r_{ii}}\left(\sum_{i=1}^n \sqrt{p_iq_i}\right)^2}{\sigma_t^2}, \tag{d}$$

and $r_{tt}\, \overline{r_{it}^2}$ is correctly substituted for $\overline{r_{ii}}$, then r_{tt} disappears entirely, and

$$\sigma_t = \overline{r_{it}}\left(\sum_{i=1}^n \sqrt{p_iq_i}\right), \tag{e}$$

*It should be pointed out that the relationship between r_{it}, r_{tt}, and r_{it} which was used in the Kuder-Richardson article is not basic to the derivation of their formulas (20) and (21). There is no implication in the present paper that those formulas need revision.

instead of their equation (10),

$$r_{tt} = \frac{\overline{r_{it}}\left(\sum_{j=1}^{n}\sqrt{p_j q_j}\right)}{\sigma_t}.$$

Actually, the amount of error incurred when (18) is used instead of (19) is quite small, if n is large. In general, this is true in the Kuder-Richardson developments described above, where a single item is correlated with the entire test. If, for example, n is taken as 100 and r_{ii} (or r_{jj}) is taken as .10, then $r_{it} = .303$ and $r_{jt} = .330$, a difference of only .027 between the spurious and the non-spurious correlations. Similarly, the error in r_{tt} is small. If the spurious value, r_{jt} , is used in (18), then r_{tt} is found to be .930 instead of .917 if (19) is (properly) used. If, on the other hand, test t is effectively not much longer than test j, then the difference in correlations can be quite appreciable. Suppose, for example, $n = 4$ and r_{ii} (or r_{jj}) = .50. Then $r_{it} = .633$ and $r_{jt} = .791$. If the spurious correlation, r_{jt} , is improperly applied in equation (18), then r_{tt} will appear to be much higher than it should—.90 instead of .80.

Tables 1 and 2 describe the results of some computations which serve to illustrate the usefulness of the estimates made in equations (6) and (12). Table 1 relates to the reliability of speeded tests, and Table 2 to power tests. In Table 1, the results are presented for four replications (with variations) of an experiment in which four randomly chosen groups of male college-level students were administered separately-timed speeded tests in mathematics. Test 1 contained 16 items in free-answer and multiple-choice form, alternately presented. Tests 2 and 3 were each composed of two tests—37 free-answer items and 37 multiple-choice items—each separately timed. Thus, scores on Test 1 were derived from 16 items, while scores on Tests 2 and 3 were each derived from 74 items; all three tests contained both free-answer and multiple-choice items.*

For each of the four groups of students, Table 1 gives three estimates of reliability of Test 2, and three for Test 3. The value r_{23} may be considered, as it has in the past, to be the reliability of either test, with the (unwarranted) presumption that the two tests are of equivalent length. In addition, r_{22} is estimated from its correlation with the shorter test, 1, adjusted for test length, and also from its correlation with Test 3, also adjusted for differences in test length. Similarly, the reliability of Test 3 is estimated from its correlation with Test 1 and also from its correlation with Test 2. It is seen that the estimates of reliability for Test 2, as derived from equation (6), are close, as are the estimates of reliability for Test 3. The largest difference, .027, is that between the two estimates for r_{33} (Group C). Even this difference may

*The author wishes to express his appreciation to Dr. L. B. Plumlee for providing the data summarized in Table 1.

TABLE 1

Reliabilities and Ratios of Effective Test Lengths for Speeded Tests: Case I

Group	N	Standard Deviations			Correlations			r_{22} Est. from		r_{33} Est. from		Ratios of Effective Lengths			
		Test 1	Test 2	Test 3	r_{12}	r_{13}	r_{23}	r_{12}	r_{23}	r_{13}	r_{23}	2:1	3:1	3:2	3:1/2:1
A	137	3.325	10.369	11.044	.767	.773	.899	.878	.896	.887	.902	3.57	3.81	1.07	1.07
B	139	2.907	10.105	10.166	.748	.761	.886	.878	.886	.885	.886	4.08	4.07	1.01	1.00
C	139	3.223	9.847	10.958	.776	.769	.909	.882	.905	.887	.914	3.47	3.92	1.12	1.13
D	139	3.263	10.770	10.274	.782	.732	.876	.892	.879	.859	.873	3.76	3.70	.95	.98

TABLE 2

Reliabilities and Ratios of Effective Test Lengths for Unspeeded Tests: Case II

Test	Group	N	Standard Deviations		r_{jt}	r_{tt}		Ratios of Effective Lengths	Ratios of Numbers of Items
			Subtest j	Total Test t		Eq. (12)	K-R (20)	t:j	t:j
VN-0	A	500	8.387	32.535	.954	.968	.977	4.07	4.17
VN-0	A	500	7.746	32.535	.958	.974	.977	4.38	4.17
VN-0	A	500	8.583	32.535	.962	.973	.977	3.94	4.05
VN-0	A	500	8.392	32.535	.964	.975	.977	4.02	4.05
VN-2	B	500	5.631	21.116	.918	.940	.942	4.08	4.17
VN-3	C	500	6.631	24.801	.929	.947	.956	4.03	4.05

be accounted for in part by the lack of complete parallelism between Tests 2 and 3 and Test 1. (It is recalled that Tests 2 and 3 had two time limits, while Test 1 had only one.)

The right-hand side of Table 1 describes the ratios of effective test lengths, as determined from equation (3). In general, Tests 2 and 3 appear to be 3.5 to 4.1 times as long as Test 1, in spite of the fact that they contain about 4.6 times as many items as Test 1. It is also observed that in Groups A and C Test 3 appears to be effectively longer than Test 2, which accounts for the slightly higher estimates of reliability for Test 3. In Group B, the lengths are seen to be about equal, while in Group D, Test 3 is the shorter of the two. Finally, in the last two columns it is observed that the two independent estimates of the ratio of effective length, Test 3 to Test 2, are extremely close.

Table 2 relates to the reliability of unspeeded tests. Three forms of a 150-item verbal and numerical reasoning test were administered—in the case of $VN-0$, to an extremely heterogeneous group of over 2000 examinees, and in the case of $VN-2$ and $VN-3$, to larger, but more homogeneous groups of examinees. In the case of each test, a sample of 500 cases was drawn at random from the parent group of examinees. For Test $VN-0$, four mutually exclusive subtests were chosen, each composed of slightly less than one-fourth of the total number of items in the parent test. Scores on the subtests were then correlated with scores on the total test, yielding the values in the column headed r_{jt} . For Tests $VN-2$ and $VN-3$, one subtest was chosen for each and correlated with its parent test. Estimates of reliability were then made in each instance in accordance with equation (12) and also in accordance with Kuder and Richardson's formula (20). It is seen that in each case the two estimates of reliability are close, differing at most by .009. It is also observed that the ratios of effective lengths determined from equation (10) are similar to the ratios of the numbers of items in the subtest and parent test.

In summary, a number of equivalent methods become available for the calculation of the reliability of a whole test without making any arbitrary assumptions of the relative lengths of the subtest and total test. The formulas presented here in the case of both speeded and power tests make use only of the general assumptions that the short test or subtest is truly a representative and parallel miniature of the long test, in terms of item content and level and spread of item difficulty. The amount by which the correlation between the short (or subtest) and long test is to be "stepped up" is determined and incorporated in the equations.

In practice, equation (6) should be used with speeded tests where the administration of an additional test is required to determine reliability. The equivalent alternative is to use equation (5) in conjunction with (3). In the case of power tests, three general procedures are possible:

(1) A double administration with mutually exclusive but parallel tests, in which reliability is to be determined from the correlation between the

parallel tests. In this case as in the case of the speeded tests, equation (6) would be used, or (5) in conjunction with (3).

(2) A single administration in which reliability is to be determined by breaking off a subtest of items parallel to the total test, and correlating the subtest score with the total test score. In this case equation (12) seems to be practicable, since it involves only one correlation and its by-product standard deviations. The equivalent alternative to (12) is equation (14), which is concise algebraically, but considerably more laborious.

(3) A single administration in which reliability is to be determined by splitting the total test into two parallel subtests (not necessarily of equal length), and correlating the subtests. In this case, equations (15) and (16) are appropriate. Formula (14) should also be mentioned here, and is appropriate, except for the reservation noted in the preceding paragraph.

REFERENCES

1. Cronbach, L. J. Test "reliability": its meaning and determination. *Psychometrika*, 1947, **12**, 1–16.
2. Cronbach, L. J., and Warrington, W. G. Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika*, 1951, **16**, 167–188.
3. Guilford, J. P. Fundamental statistics in psychology and education (2nd ed.). New York: McGraw-Hill Book Co., 1950.
4. Gulliksen, H. The reliability of speeded tests. *Psychometrika*, 1950, **15**, 259–269.
5. Guttman, L. A basis for estimating test-retest reliability. *Psychometrika*, 1945, **10**, 255–282.
6. Kuder, G. F., and Richardson, M. W. The theory and estimation of test reliability. *Psychometrika*, 1937, **2**, 151–160.
7. Thorndike, R. L. Reliability. In Lindquist, E. F., Educational measurement. Washington, D. C.: American Council on Education, 1951.

Manuscript received 5/19/52

Revised manuscript received 6/28/52