PSYCHOMETRIKA—VOL. 16, NO. 3
SEPTEMBER, 1951

COEFFICIENT ALPHA AND THE INTERNAL STRUCTURE OF TESTS*

LEE J. CRONBACH
UNIVERSITY OF ILLINOIS

A general formula (α) of which a special case is the Kuder-Richardson coefficient of equivalence is shown to be the mean of all split-half coefficients resulting from different splittings of a test. α is therefore an estimate of the correlation between two random samples of items from a universe of items like those in the test. α is found to be an appropriate index of equivalence and, except for very short tests, of the first-factor concentration in the test. Tests divisible into distinct subtests should be so divided before using the formula. The index $\bar{r}_{ij}$, derived from α, is shown to be an index of inter-item homogeneity. Comparison is made to the Guttman and Loevinger approaches. Parallel split coefficients are shown to be unnecessary for tests of common types. In designing tests, maximum interpretability of scores is obtained by increasing the first-factor concentration in any separately-scored subtest and avoiding substantial group-factor clusters within a subtest. Scalability is not a requisite.

I. *Historical Resumé*

Any research based on measurement must be concerned with the accuracy or dependability or, as we usually call it, reliability of measurement. A reliability coefficient demonstrates whether the test designer was correct in expecting a certain collection of items to yield interpretable statements about individual differences (25).

Even those investigators who regard reliability as a pale shadow of the more vital matter of validity cannot avoid considering the reliability of their measures. No validity coefficient and no factor analysis can be interpreted without some appropriate estimate of the magnitude of the error of measurement. The preferred way to find out how accurate one's measures are is to make two independent measurements and compare them. In practice, psychologists and educators have often not had the opportunity to recapture their subjects for a second test. Clinical tests, or those used for vocational guidance, are generally worked into a crowded schedule, and there is always a de-

*The assistance of Dora Damrin and Willard Warrington is gratefully acknowledged. Miss Damrin took major responsibility for the empirical studies reported. This research was supported by the Bureau of Research and Service, College of Education.

sire to give additional tests if any extra time becomes available. Purely scientific investigations fare little better. It is hard enough to schedule twenty tests for a factorial study, let alone scheduling another twenty just to determine reliability.

This difficulty was first circumvented by the invention of the split-half approach, whereby the test is rescored, half the items at a time, to get two estimates. The Spearman-Brown formula is then applied to get a coefficient similar to the correlation between two forms. The split-half Spearman-Brown procedure has been a standard method of test analysis for forty years. Alternative formulas have been developed, some of which have advantages over the original. In the course of our development, we shall review those formulas and show relations between them.

The conventional split-half approach has been repeatedly criticized. One line of criticism has been that split-half coefficients do not give the same information as the correlation between two forms given at different times. This difficulty is purely semantic (9, 14) ; the two coefficients are measures of different qualities and should not be identified by the same unqualified appellation "reliability." A retest after an interval, using the identical test, indicates how stable scores are and therefore can be called a coefficient of *stability*. The correlation between two forms given virtually at the same time, is a coefficient of *equivalence*, showing how nearly two measures of the same general trait agree. Then the coefficient using comparable forms with an interval between testings is a coefficient of equivalence and stability. This paper will concentrate on coefficients of equivalence.

The split-half approach was criticized, first by Brownell (3), later by Kuder and Richardson (26), because of its lack of uniqueness. Instead of giving a single coefficient for the test, the procedure gives different coefficients depending on which items are grouped when the test is split in two parts. If one split may give a higher coefficient than another, one can have little faith in whatever result is obtained from a single split. This criticism is with equal justice applicable to any equivalent-forms coefficient. Such a coefficient is a property of a pair of tests, not a single test. Where four forms of a test have been prepared and intercorrelated, six values are obtained, and no one of these is *the* unique coefficient for Form A; rather, each is the coefficient showing the equivalence of one form to another specific form.

Kuder and Richardson derive a series of coefficients using data from a single trial, each of them being an approximation to the interform coefficient of equivalence. Of the several formulas, one has been justifiably preferred by test workers. In this paper we shall be especially concerned with this, their formula (20):

$$r_{tt(KR20)} = \frac{n}{n-1}\left(1 - \frac{\sum\limits_{i} p_i q_i}{\sigma_t^2}\right); \underline{\ (i = 1, 2, \cdots n)}. \tag{1}$$

Here, i represents an item, p_i the proportion receiving a score of 1, and q_i the proportion receiving a score of zero on the item.

We can write the more general formula

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum\limits_{i} V_i}{V_t}\right). \tag{2}$$

Here V_t is the variance of test scores, and V_i is the variance of item scores after weighting. This formula reduces to (1) when all items are scored 1 or zero. The variants reported by Dressel (10) for certain weighted scorings, such as Rights-minus-Wrongs, are also special cases of (2), but for most data computation directly from (2) is simpler than by Dressel's method. Hoyt's derivation (20) arrives at a formula identical to (2), although he draws attention to its application only to the case where items are scored 1 or 0. Following the pattern of any of the other published derivations of (1) (19, 22), making the same assumptions but imposing no limit on the scoring pattern, will permit one to derive (2).

Since each writer offering a derivation used his own set of assumptions, and in some cases criticized those used by his predecessors, the precise meaning of the formula became obscured. The original derivation unquestionably made much more stringent assumptions than necessary, which made it seem as if the formula could properly be applied only to rare tests which happened to fit these conditions. It has generally been stated that α gives a lower bound to "the true reliability"—whatever that means to that particular writer. In this paper, we take formula (2) as given, and make no assumptions regarding it. Instead, we proceed in the opposite direction, examining the properties of α and thereby arriving at an interpretation.

We introduce the symbol α partly as a convenience. "Kuder-Richardson Formula 20" is an awkward handle for a tool that we expect to become increasingly prominent in the test literature. A second reason for the symbol is that α is one of a set of six analogous coefficients (to be designated β, γ, δ, etc.) which deal with such other concepts as like-mindedness of persons, stability of scores, etc. Since we are concentrating in this paper on equivalence, the first of the six properties, description of the five analogous coefficients is reserved for later publication.

Critical comments on the Kuder-Richardson formula have been primarily directed to the fact that when inequalities are used in deriving a lower bound, there is no way of knowing whether a particular coefficient is a close estimate of the desired measure of equivalence or a gross underestimate. The Kuder-Richardson method is an overall measure of internal consistency, but a test which is not internally homogeneous may nonetheless have a high correlation with a carefully-planned equivalent form. In fact, items within each test may correlate zero, and yet the two tests may correlate perfectly if there is item-to-item correspondence of content.

The essential problem set in this paper is: How shall α be interpreted? α, we find, is the average of all the possible split-half coefficients for a given test. Juxtaposed with further analysis of the variation of split-half coefficients from split to split, and with an examination of the relation of α to item homogeneity, this relation leads to recommendations for estimating coefficients of equivalence and homogeneity.

II. *A Comparison of Split-Half Formulas*

The problem set by those who have worked out formulas for split-half coefficients is to predict the correlation between two equivalent whole tests, when data on two half-tests are at hand. This requires them to define equivalent tests in mathematical terms.

The first definition is that introduced by Brown (2) and by Spearman (33), namely, that we seek to predict correlation with a test whose halves are c and d, possessing data from a test whose halves are a and b, and that

$$\begin{gathered} V_a = V_b = V_c = V_d; \quad \text{and} \\ r_{ab} = r_{ac} = r_{ad} = r_{bc} = r_{bd} = r_{cd}\,. \end{gathered} \tag{3}$$

This assumption or definition is far from general. For many splittings $V_a \neq V_b$, and an equivalent form conforming to this definition is impossible.

A more general specification of equivalence credited to Flanagan [see (25)] is that

$$\begin{gathered} V_{(a+b)} = V_{(c+d)}; \quad \text{and} \\ r_{ab}\sigma_a\sigma_b = r_{ad}\sigma_a\sigma_d = r_{bc}\sigma_b\sigma_c = r_{cd}\sigma_c\sigma_d = \cdots\,. \end{gathered} \tag{4}$$

This assumption leads to various formulas which are collected in the first column of Table 1. All formulas in Column A are mathematically identical and interchangeable.

TABLE 1
Formulas for Split-Half Coefficients

Entering Data*		Formulas Assuming Equal Covariances Between Half-Tests		Formulas Assuming $\sigma_a = \sigma_b$
$r_{ab}\ \sigma_a\ \sigma_b$	1A†	$\dfrac{4\sigma_a\sigma_b r_{ab}}{\sigma_a^2 + \sigma_b^2 + 2\sigma_a\sigma_b r_{ab}}$	1B‡	$\dfrac{2r_{ab}}{1 + r_{ab}}$
$\sigma_t\ \sigma_a\ \sigma_b$	2A§	$2\left(1 - \dfrac{\sigma_a^2 + \sigma_b^2}{\sigma_t^2}\right)$		
$\sigma_t\ \sigma_a\ r_{at}$	3A‖	$\dfrac{4(r_{at}\sigma_a\sigma_t - \sigma_a^2)}{\sigma_t^2}$		
$\sigma_t\ \sigma_d$	4A¶	$1 - \dfrac{\sigma_d^2}{\sigma_t^2}$	4B (≡4A)	$1 - \dfrac{\sigma_d^2}{\sigma_t^2}$
$\sigma_a\ \sigma_d\ r_{ad}$	5A	$\dfrac{4(\sigma_a^2 - \sigma_a\sigma_d r_{ad})}{4\sigma_a^2 + \sigma_d^2 - 4\sigma_a\sigma_d r_{ad}}$	5B	$\dfrac{2(2\sigma_a^2 - \sigma_d^2)}{4\sigma_a^2 - \sigma_d^2}$

*In this table, a and b are the half-test scores. $t = a+b$, $d = a-b$.

†After Flanagan (25)

‡Spearman-Brown (2, 33)

§Guttman (19)

‖After Mosier (28)

¶Rulon (31)

When a particular split is such that $\sigma_a = \sigma_b$, the Flanagan requirement reduces to the original Spearman-Brown assumption, and in that case we arrive at the formulas in Column B. Formulas 1B and 5B are not identical, since the assumption enters the formulas in different ways. No short formula is provided opposite 2A or 3A, since these exact formulas are themselves quite simple to compute.

Because of the wide usage of Formula 1B, the Spearman-Brown, it is of interest to determine how much difference it makes which assumption is employed. If we divide 1B by any of the formulas in Column A we obtain the ratio

$$k_1 = \frac{2mr + m^2 + 1}{2m(1 + r)} = \frac{1}{(1 + r)}\left(\frac{1 + m^2 + r}{2m}\right), \qquad (5)$$

in which $\underline{m} = \sigma_b/\sigma_a$, $\sigma_a < \sigma_b$, and r signifies r_{ab}. The ratio when 5B is divided by any of the formulas in the first column is as follows:

$$k_5 = \frac{(2mr - m^2 + 1)(1 + 2mr + m^2)}{2mr(2mr - m^2 + 3)}. \tag{6}$$

When m equals 1, that is, when the two standard deviations are equal, the formula in Column B is identical to that in Column A. As Table 2 shows, there is increasing disagreement between Formula 1B and those in Column A as m departs from unity. The estimate by the Spearman-Brown formula is always slightly larger than the coefficient of equivalence computed by the more tenable definition of comparability.

TABLE 2

Ratio of Spearman-Brown Estimate to More Exact Split-Half Estimate of Coefficient of Equivalence when S.D.'s are Unequal

Ratio of Half-Test S.D.'s (greater/lesser)	Correlation Between Half-Tests					
	.00	.20	.40	.60	.80	1.00
1	1	1	1	1	1	1
1.1	1.005	1.004	1.003	1.003	1.003	1.002
1.2	1.017	1.014	1.012	1.010	1.009	1.008
1.3	1.035	1.029	1.025	1.022	1.020	1.017
1.4	1.057	1.048	1.041	1.036	1.032	1.029
1.5	1.083	1.069	1.060	1.052	1.046	1.042

Formula 5B is not so close an approximation to the results from formulas in Column A. When m is 1.1, for example, the values of k_5 are as follows: for $r = .20$, .62; for $r = .60$, .70; for $r = 1.00$, .999.

It is recommended that the interchangeable formulas 2A and 4A be used in obtaining split-half coefficients. These formulas involve no assumptions contradictory to the data. They are therefore preferable to the Spearman-Brown formula. However, if the ratio of the standard deviations of the half-tests is between .9 and 1.1, the Spearman-Brown formula gives essentially the same result. This finding agrees with Kelley's earlier analysis of much the same question (2, 3).

III. *α as the Mean of Split-Half Coefficients*

To demonstrate the relation between α and the split-half formulas, we shall need the following notation:

Let n be the number of items.

The test t is divided into two half-tests, a and b. i' will designate any item of half-test a, and i'' will designate any item of half-test b. Each half-test contains n' items, where $n' = n/2$.

V_t, V_a, and V_b are the variances of the total test and the respective half-tests.

C_{ij} is the covariance of two items i and j.

C_a is the total covariance for all items in pairs within half-test a, each pair counted once; C_b is the corresponding "within-test" covariance for b.

C_t is the total covariance of all item pairs within the test.

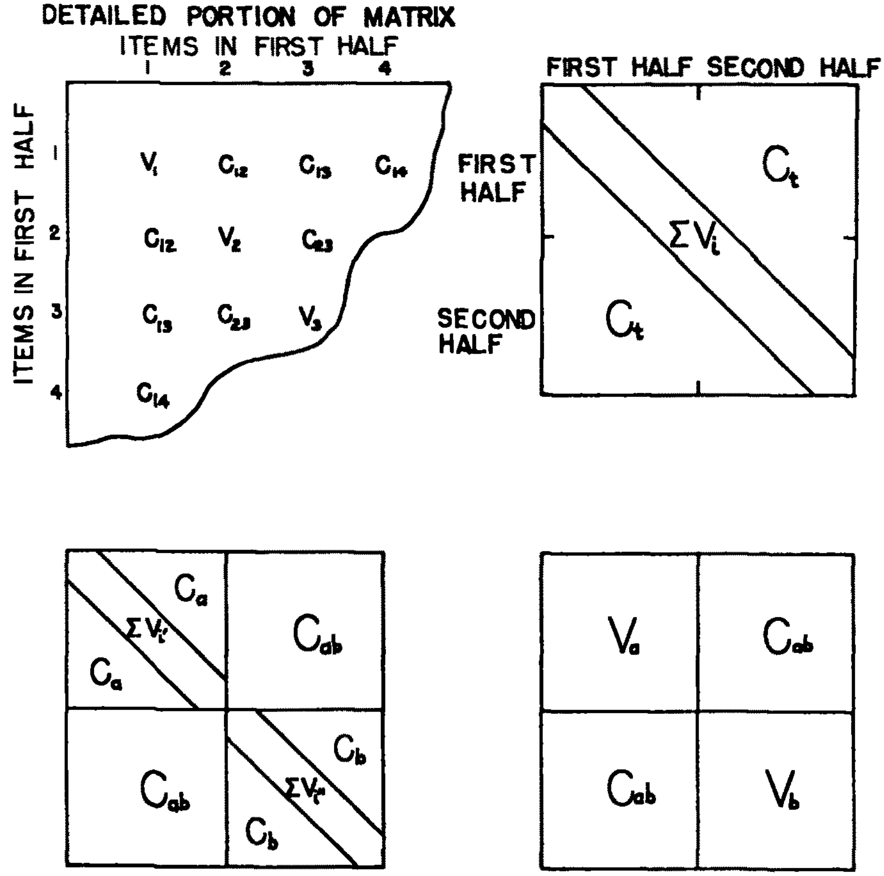

FIGURE 1

Schematic Division of the Matrix of Item Variances and Covariances.

C_{ab} is the total covariance of all item pairs such that one item is within a and the other is within b; it is the "between halves" covariance.

Then

$$C_{ab} = r_{ab}\sigma_a\sigma_b; \tag{7}$$

$$C_t = C_a + C_b + C_{ab}; \tag{8}$$

$$V_t = V_a + V_b + 2C_{ab} = \sum_i V_i + 2C_t; \text{ and} \tag{9}$$

$$V_a = \sum_{i'} V_{i'} + 2C_a \text{ and } V_b = \sum_{i''} V_{i''} + 2C_b. \tag{10}$$

These identities are readily visible in the sketches of Figure 1, which is based on the matrix of item covariances and variances. Each point along the diagonal represents a variance. The sum of all entries in the square is the test variance.

Rewriting split-half formula 2A, we have

$$r_{tt} = 2\left(1 - \frac{V_a + V_b}{V_t}\right) = 2\,\frac{V_t - V_a - V_b}{V_t}. \tag{11}$$

$$r_{tt} = \frac{4C_{ab}}{V_t}. \tag{12}$$

This indicates that whether a particular split gives a high or low coefficient depends on whether the high interitem covariances are placed in the "between halves" covariance or whether the items having high correlations are placed instead within the same half.

Now we rewrite α:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right) = \frac{n}{n-1}\left(\frac{V_t - \sum_i V_i}{V_t}\right). \tag{13}$$

$$\alpha = \frac{n}{n-1} \cdot \frac{2C_t}{V_t}. \tag{14}$$

$$\bar{C}_{ij} = \frac{C_t}{n(n-1)/2}. \tag{15}$$

Therefore

$$\alpha = \frac{n^2\,\bar{C}_{ij}}{V_t}. \tag{16}$$

We proceed now by determining the mean coefficient from all $(2n')!/2(n'!)^2$ possible splits of the test. From (12),

$$\bar{r}_{tt} = \frac{4\,\bar{C}_{ab}}{V_t}. \qquad (17)$$

In any split, a particular C_{ij} has a probability of $\frac{n}{2(n-1)}$ of falling into the between-halves covariance C_{ab}. Then over all splits,

$$\sum C_{ab} = \frac{(2n')!}{2(n'!)^2}\,\frac{n}{2(n-1)}\sum_i\sum_j C_{ij};\ (i=1,2,\cdots n-1;\ \underline{j=i+1},\cdots;n). \qquad (18)$$

But

$$\sum_i\sum_j C_{ij} = \frac{n(n-1)}{2}\bar{C}_{ij}. \qquad (19)$$

$$\sum C_{ab} = \frac{(2n')!}{2(n'!)^2}\,\frac{n^2}{4}\bar{C}_{ij}, \qquad (20)$$

and

$$\bar{C}_{ab} = \frac{n^2}{4}\bar{C}_{ij}. \qquad (21)$$

From (17),

$$\bar{r}_{tt} = \frac{4n^2}{4V_t}\bar{C}_{ij} = \frac{n^2\,\bar{C}_{ij}}{V_t}. \qquad (22)$$

Therefore

$$\bar{r}_{tt} = a\,. \qquad (23)$$

From (14), we can also write a in the form

$$a = \frac{n}{n-1}\,\frac{\sum_i\sum_j C_{ij}}{V_t};\ (i,j=1,2,\cdots n;\ i\neq j). \qquad (24)$$

This important relation states a clear meaning for a as $n/(n-1)$ times the *ratio of interitem covariance to total variance.* The multiplier $n/(n-1)$ allows for the proportion of variance in any item which is due to the same elements as the covariance.

α as a special case of the split-half coefficient. Not only is α a function of all the split-half coefficients for a test; it can also be shown to be a special case of the split-half coefficient.

If we assume that the test is divided into equivalent halves such that $\overline{C}_{i'i''}$ (i.e., C_{ab}/n'^2) equals $\overline{C}_{ij}$, the assumptions for formula 2A still hold. We may designate the split-half coefficient for this splitting as r_{tt_0}.

$$r_{tt} = \frac{4\,C_{ab}}{V_t}. \tag{12}$$

Then

$$r_{tt_0} = \frac{4n'^2\,\overline{C}_{i'i''}}{V_t} = \frac{4n'^2\,\overline{C}_{ij}}{V_t} = \frac{n^2\,\overline{C}_{ij}}{V_t}. \tag{25}$$

From (16),

$$r_{tt_0} = \alpha. \tag{26}$$

This amounts to a proof that α is an exact determination of the parallel-form correlation when we can assume that the mean covariance between parallel items equals the mean covariance between unpaired items. This is the least restrictive assumption usable in "proving" the Kuder-Richardson formula.

α as the equivalence of random samples of items. The foregoing demonstrations show that α measures essentially the same thing as the split-half coefficient. If all the splits for a test were made, the mean of the coefficients obtained would be α. When we make only one split, and make that split at random, we obtain a value somewhere in the distribution of which α is the mean. If split-half coefficients are distributed more or less symmetrically, an obtained split-half coefficient will be higher than α about as often as it is lower than α. This average that is α is based on the very best splits and also on some very poor splits where the items going into the two halves are quite unlike each other.

Suppose we have a universe of items for which the mean covariance is the same as the mean covariance within the given test. Then suppose two tests are made by twice sampling n items at random from this universe without replacement, and administered at the same sitting. Their correlation would be a coefficient of equivalence. The mean of such coefficients would be the same as the computed α. α is therefore an estimate of the correlation expected between two tests drawn at random from a pool of items like the items in this test. Items are not selected at random for psychological tests where any differentiation among the items' contents or difficulties permits a planned selection. Two planned samplings may be expected to have higher correlations than two random samplings, as Kelley pointed out (25). We shall show that this difference is usually small.

IV. *An Examination of Previous Interpretations and Criticisms of* α

1. *Is* α *a conservative estimate of reliability?* The findings just presented call into question the frequently repeated statement that α is a conservative estimate or an underestimate or a lower bound to "the reliability coefficient." The source of this conception is the original derivation, where Kuder and Richardson set up a definition of two equivalent tests, expressed their correlation algebraically, and proceeded to show by inequalities that α was lower than this correlation. Kuder and Richardson assumed that corresponding items in test and parallel test have the same common content and the same specific content, i.e., that they are as alike as two trials of the same item would be. In other words, they took the zero-interval retest correlation as their standard. Guttman also began his derivation by defining equivalent tests as identical. Coombs (6) offers the somewhat more satisfactory name "coefficient of precision" for this index which reports the absolute minimum error to be found if the same instrument is applied twice independently to the same subject. A coefficient of stability can be obtained by making the two observations with any desired interval between. A rigorous definition of the coefficient of precision, then, is that it is the limit of the coefficient of stability, as the time between testings becomes infinitesimal.

Obviously, any coefficient of equivalence is less than the coefficient of precision, for one is based on a comparison of different items, the other on two trials of the same items. To put it another way: α or any other coefficient of equivalence treats the specific content of an item as error, but the coefficient of precision treats it as part of the thing being measured. It is very doubtful if testers have any practical need for a coefficient of precision. There is no practical testing problem where the items in the test and only these items constitute the trait under examination. We may be unable to compose more items because of our limited skill as testmakers but any group of items in a test of intelligence or knowledge or emotionality is regarded as a sample of items. If there weren't "plenty more where these came from," performance on the test would not represent performance on any more significant variable.

We therefore turn to the question, does α underestimate appropriate coefficients of equivalence? Following Kelley's argument, the way to make equivalent tests is to make them as similar as possible, similar in distribution of item difficulty and in item content. A pair of tests so designed that corresponding items measure the same factors, even if each one also contains some specific variance, will have a higher correlation than a pair of tests drawn at random from the pool of items. A planned split, where items in opposite halves are as similar as the test permits, may logically be expected to have a higher between-halves covariance than within-halves covariance, and in that case, the obtained coefficient would be larger than α. α is the same type of coefficient as the split-half coefficient, and while it may be lower, it may also be higher than the value obtained by actually splitting a particular test at random. Both the random or odd-even split-half coefficient and α will theoretically be lower than the coefficient from parallel forms or parallel splits.

2. *Is α less than the coefficient of stability?* Some writers expect α to be lower than the coefficient of stability. Thus Guttman says (34, p. 311):

> For the case of scale scores, then, . . . we have the assurance that if the items are approximately scalable [in which case α will be high], then they necessarily have very substantial test-retest reliability.

Guilford says (16, p. 485):

> There can be very low internal consistency and yet substantial or high retest reliability. It is probably not true, however, that there can be high internal consistency and at the same time low retest reliability, except after very long time intervals. If the two indices of reliability disagree for a test, we can place some confidence in the inference that the test is heterogeneous.

The comment by Guttman is based on sound thinking, provided we reinterpret test-retest coefficient on the basis of the context of the comment to refer to the instantaneous retest (i.e., coefficient of precision) rather than the retest after elapsed time. Guilford's statement is acceptable only if viewed as a summary of his experience. There is no mathematical necessity for his remarks to be true. In the coefficient of stability, variance in total score between trials (within persons) is regarded as a source of error, and variance in specific factors (between items within persons) within trials is regarded as true variance. In the coefficient of equivalence, such as α, this is just reversed: variance in specific factors is treated as error. Variation between trials is non-existent and does not reduce true variance (9). Whether the coefficient of stability is higher or lower than the coefficient of equivalence depends on the relative magnitude of these variances, both of which are likely to be small for long tests of stable variables. Tests are also used for unstable variables such as mood, morale, social interaction, and daily work output, and studies of this sort are becoming increasingly prominent. Suppose one builds a homogeneous scale to obtain students' evaluations of each day's classwork, the students marking the checklist at the end of each class hour. Homogeneous items could be found for this. Yet the scale would have marked instability from day to day, if class activities varied or the topics discussed had different interest value for different students.

The only proper conclusion is that α may be either higher or lower than the coefficient of stability over an interval of time.

3. *Are coefficients from parallel splits appreciably higher than random-split coefficients or α?* The logical presumption is strong that planned splits as proposed by Kelley (25) and Cronbach (7) would yield coefficients nearer to the equivalent-tests coefficient than random splits do. There is still the empirical question whether this advantage is large enough to be considered seriously. This raises two questions: Is there appreciable variation in coefficients from split to split? If so, does the judgment made in splitting the test into *a priori* equivalent halves raise the coefficient? Brownell (3), Cronbach (8), and Clark (5) have compared coefficients obtained by splitting a test in many ways. There is doubt that the variation among coefficients is ordinarily a serious matter; Clark in particular found that variation from split to split was small compared to variation arising from sampling of subjects.

Empirical evidence. To obtain further data on this question, two analyses were made. One employs responses of 250 ninth-grade boys who took Mechanical Reasoning Test Form A of the Differential Abilities Tests. The second study uses a ten-item morale scale, adapted from the Rundquist-Sletto General Morale Scale by Donald M. Sharpe and administered by him to teachers and school administrators.*

The Mechanical Reasoning Test seems to contain items requiring specific knowledges regarding pulleys, gears, etc. Other items seem to be answerable on the basis of general experience or reasoning. The items seemed to represent sufficiently heterogeneous content that grouping into parallel splits would be possible. We found, however, that items grouped on *a priori* grounds had no higher correlations than items believed to be unlike in content. This finding is con-

*Thanks are expressed to Dr. A. G. Wesman and the Psychological Corporation, and to Dr. Sharpe, for making available the data for the two studies, respectively.

firmed by Air Force psychologists who made a similar attempt to categorize items from a mechanical reasoning test and found that they could not. These items, they note, "are typically complex factorially" (15, p. 309).

Eight items which some students omitted were dropped. An item analysis was made for 50 papers. Using this information, ten parallel splits were made such that items in opposite halves had comparable difficulty. These we call Type I splits. Then eight more splits were made, placing items in opposite halves on the basis of both difficulty and apparent content (Type II splits). Fifteen random splits were made. For all splits, Formula 2A was applied, using the 200 remaining cases. Results appear in Table 3.

TABLE 3

Summary of Data from Repeated Splittings of Mechanical Reasoning Test (60 items; $\alpha = .811$)

	All Splits			Splits Where $1.05 > \sigma_b/\sigma_a > .95$		
Type of Split	No. of Coefficients	Range	Mean	No. of Coefficients	Range	Mean
Random	15	.779–.860	.810	8	.795–.860	.817
Parallel Type I	10	.798–.846	.820	6	.798–.846	.822
Parallel Type II	8	.801–.833	.817	4	.809–.826	.818

There are only 126 possible splits for the morale test, and it is possible to compute all half-test standard deviations directly from the item variances and covariances. Of the 126 splits, six were designated in advance as Type II parallel splits, on the basis of content and an item analysis of a supplementary sample of papers. Results based on 200 cases appear in Table 4.

TABLE 4

Summary of Data from Repeated Splittings of Morale Scale (10 items; $\alpha = .715$)

	All Splits			Splits Where $1.1 > \sigma_b/\sigma_a > .9$		
Type of Split	No. of Coefficients	Range	Mean	No. of Coefficients	Range	Mean
All Splits	126	.609–.797	.715	82	.609–.797	.717
Parallel (Type II)	6	.681–.780	.737	5	.712–.780	.748

The highest and lowest coefficients for the mechanical test differ by only .08, a difference which would be important only when a very precise estimate of reliability is needed. The range for the morale scale is greater (.20), but the probability of obtaining one of the extreme values in sampling is slight. Our findings agree with Clark, that the variation from split to split is less than the variation expected from sample to sample for the same split. The standard error of a Spearman-Brown coefficient based on 200 cases using the same split is .03 when $r_{tt} = .8$, .04 when $r_{tt} = .7$. The former value compares with a standard deviation of .02 for all random-split coefficients of the mechanical test. The standard error of .04 compares with a standard deviation of .035 for the 126 coefficients of the morale test.

This bears on Kelley's comment on proposals to obtain a unique estimate: "A determinate answer would result if the mean for all possible spilts were gotten, but, even neglecting the labor involved, this would seem to contravene the judgment of comparability." (25, p. 79). As our tables show, the splittings where half-test standard deviations are unequal, which "contravene the judgment of comparability," have coefficients about like those which have equal standard deviations.

Combining our findings with those of Clark and Cronbach we have studies of seven tests which seem to show that the variation from split to split is too small to be of practical importance. Brownell finds appreciable variation, however, for the four tests he studied. The apparent contradiction is explained by the fact that the former results applied to tests having fairly large coefficients of equivalence (.70 or over). Brownell worked with tests whose coefficients were much lower, and the larger range of r's does not represent any greater variation in z values at this lower level.

In Tables 3 and 4, the values obtained from deliberately equated half-tests differ slightly, but only slightly, from those for random splits. Where α is .715 for the morale scale, the mean of parallel splits is .748—a difference of no practical importance. One parallel split reaches .780, but this split could not have been defended *a priori* as more logical than the other planned splits. In Table 3, we find that neither Type I nor Type II splits averaged more than .01 higher than α. Here, then, is evidence that the sort of judgment a tester might make on typical items, knowing their content and difficulty, does not, contrary to the earlier opinion of Kelley and Cronbach, permit him to make more comparable half-tests than would be obtained by random splitting. The data from Cronbach's earlier study agree with this. This conclusion seems to apply to tests of any length (the morale scale has only ten items). Where items fall into obviously diverse subgroups in either content or difficulty, as, say, in the California Test of Mental Maturity, the tester's judgment could provide a better-than-random split. It is dubious whether he could improve on a random division *within subtests.*

It should be noted that in this empirical study no attempt was made to divide items on the basis of r_{it}, as Gulliksen (18, p. 207-210) has recently suggested. Provided this is done on a large sample of cases other than those used to estimate r_{tt}, Gulliksen's plan might indeed give parallel-split coefficients which are consistently at least a few points higher than α.

The failure of the data to support our expectation led to a further study of the problem. We discovered that even tests which seem to be heterogeneous are often highly saturated with the first factor among the items. This forces us not only to extend the interpretation of α, but also to reexamine certain theories of test design.

Factorial composition of the test variance. To make fully clear the relations involved, our analytic procedure will be spelled out in detail. We postulate that the variance of any item can be divided among $k + 1$ orthogonal factors (k common with other items and one unique). Of these, we shall refer to the first, f_1, as the general factor, even though it is possible that some items would have a zero loading on this factor.* Then if f_{zi} is the loading of common factor z on item i,

$$1.00 = N^2 (f^2_{1i} + f^2_{2i} + f^2_{3i} + \cdots + f^2_{U_i}) . \tag{27}$$

$$C_{ij} = N^2 \sigma_i \sigma_j (f_{1i} f_{1j} + f_{2i} f_{2j} + \cdots + f_{ki} f_{kj}). \tag{28}$$

$$C_t = \Sigma\Sigma C_{ij} = N^2 \sum_i \sum_j \sigma_i \sigma_j f_{1i} f_{1j} + \cdots + N^2 \sum_i \sum_j \sigma_i \sigma_j f_{ki} f_{kj} ;$$

$$(i = 1 , 2 , \cdots n-1 ; j = i + 1 , \cdots , n). \tag{29}$$

$$V_t = N^2 \sum_i \sigma^2_i (f^2_{1i} + \cdots f^2_{ki} + f^2_{U_i}) + 2N^2 \sum_i \sum_j \sigma_i \sigma_j f_{1i} f_{1j}$$

$$+ \cdots + 2N^2 \sum_i \sum_j \sigma_i \sigma_j f_{ki} f_{kj} . \tag{30}$$

If n_1 items contain non-zero loadings on factor 1, and n_2 items contain factor 2, etc., then V_t consists of

*This factor may be a so-called primary or reference factor like Verbal, but it is more likely to be a composite of several such elements which contribute to every item.

n_1^2 terms of the form $N^2\sigma_i\sigma_j f_{1i} f_{1j}$, plus
n_2^2 terms of the form $N^2\sigma_i\sigma_j f_{2i} f_{2j}$, plus (31)
n_3^2 terms of the form $N^2\sigma_i\sigma_j f_{3i} f_{3j}$, plus and so on to
n_k^2 terms of the form $N^2\sigma_i\sigma_j f_{ki} f_{kj}$, plus
n terms of the form $N^2\sigma_i^2 f_{U_i}^2$.

We rarely know the values of the factor loadings for an actual test, but we can substitute values representing different kinds of test structure in (30) and observe the proportionate influence of each factor in the total test.

First we shall examine a test made up of a general factor and five group factors, in effect a test which might be arranged into five correlated subtests. $k = 6$. Let $n_1 = n$, so f_1 is truly general, and let $n_2 = n_3 = n_4 = n_5 = n_6 = 1/5\ n$. To keep the illustration simple, we shall assume that all items have equal variances and that any factor has the same loading (f_z) in all items where it appears. Then

$$\frac{1}{N^2\sigma_i^2} V_t = n^2 f_1^2 + \frac{n^2}{25} f_2^2 + \frac{n^2}{25} f_3^2 + \cdots + \frac{n^2}{25} f_6^2 + \sum_i f_{U_i}^2. \tag{32}$$

It follows that in this particular example, there are n^2 general factor terms, $n^2/5$ group factor terms, and only n unique factor terms. There are, in all, $6n^2/5 + n$ terms in the variance. Let f^2_{zt} be the proportion of test variance due to each factor. Then if we assume that all the terms making up the variance are of the same approximate magnitude,

$$f^2_{1t} = \frac{5n^2}{6n^2 + 5n} = \frac{5n}{6n + 5}. \tag{33}$$

$$\lim_{n \to \infty} f^2_{it} = \frac{5}{6} = .83. \tag{34}$$

$$f^2_{2t} = \cdots = f^2_{6t} = \frac{n^2/5}{6n^2 + 5n}. \tag{35}$$

$$\lim_{n \to \infty} f^2_{2t} = .03. \tag{36}$$

$$\sum_i f^2_{U_i t} = \frac{5}{6n + 5}. \tag{37}$$

$$\lim_{n \to \infty} \sum_i f^2_{U_i t} = 0. \tag{38}$$

Note that among the terms making up the variance of any test, the number of terms representing the general factor is n times the number representing item specific and error factors.

We have seen that the general factor cumulates a very large influence in the test. This is made even clearer by Figure 2, where we plot the trend in variance for a particular case of the above test structure. Here we set $k = 6$, $n_1 = n$, $n_2 = n_3 = n_4 = n_5 = n_6 = n/5$. Then we assume that each item has the composition: 9% general factor, 9% from some one group factor, 82% unique. Further, the unique variance is divided by 70/12 between error and specific stable variance. It is seen that even with unreliable items such as these, which intercorrelate only .09 or .18, the general factor quickly becomes the predominant portion of the variance. In the limit, as n becomes indefinitely large, the general factor is 5/6 of the variance, and each group factor is 1/30 of the total variance.

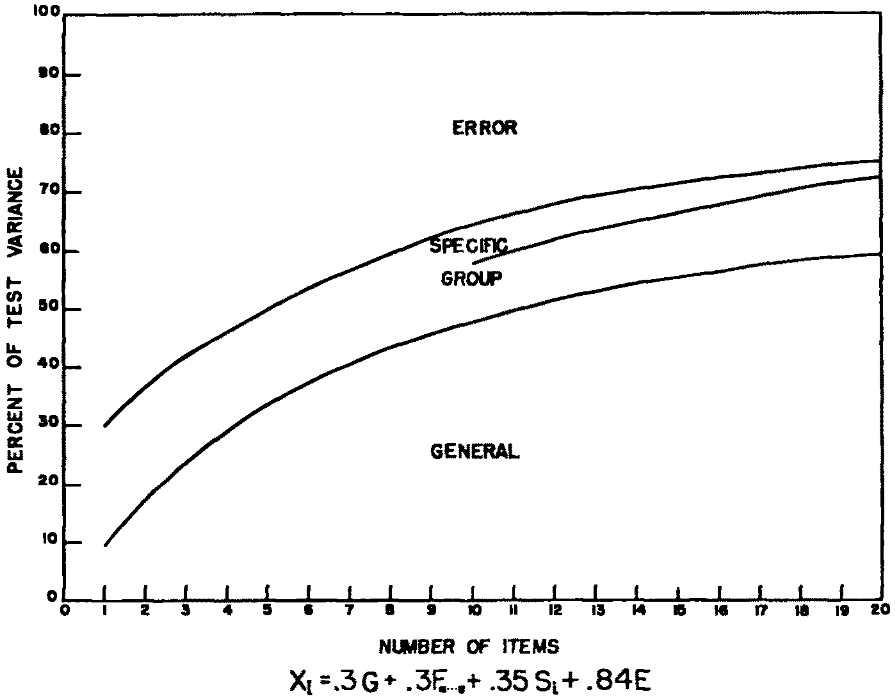

FIGURE 2

Change in Proportion of Test Variance due to General, Group, and Unique Factors among the Items as n Increases.

This relation has such important consequences that we work out two more illustrative substitutions in Table 5. We first consider the test which is very heterogeneous in one sense, in that each group of five items introduces a different group factor. No factor save factor 1 is found in more than 5 items. Here great weight in each item is given to the group factor, yet even so, the general factor quickly cumulates in the covariance terms and outweighs the group factors.

The other illustration involves a case where the general factor is much less important in the items than two group factors, each present in half the items. In this type of test, the general factor takes on some weight through cumulation, but the group factors do not fade into insignificance as before. We can generalize that when the proportion of items containing each common factor remains constant as a test is lengthened (factor loadings being constant also), the ratio of the variances contributed by any two common factors remains constant. That is, in such a test pattern each item accounts for a nearly constant fraction of the non-unique variance.

While our description has discussed number of terms, and has simplified by holding constant both item variances and factor loadings, the same general trends hold if these conditions are not imposed. The mathematical notation required is intricate, and we have not attempted a formal derivation of these general principles:

If the magnitude of item intercorrelations is the same, on the average, in successive groups of items as a test is lengthened,

(a) Specific factors and unreliability of responses on single items account for a rapidly decreasing proportion of the variance if the added items represent the same factors as the original items. Roughly, the contribution is inversely proportional to test length.

(b) The ratio in which the remaining variance is divided among the general factor and group factors

(i) is constant if these factors are represented in the added items to the same extent as in the original items;*

(ii) increases, if the group factors present in the original items have less weight in the added items.

As a test is lengthened, the general factor accounts for a larger and larger proportion of the total variance. In the case where only a few group factors are present no matter how many items are added,

*This is the case discussed in the recent paper of Guilford and Michael (17). Our conclusion is identical to theirs.

TABLE 5
Factor Composition of Tests Having Certain Item Characteristics as a Function of Test Length

Pattern	Factors	Per Cent of Variance in Any Item	Number of Items Containing Factor	Per Cent of Total Test Variance (assuming equal item variances)				
				$n=1$	$n=5$	$n=25$	$n=100$	$n\to\infty$
One general factor,	f_1	9	n	9	13	44	76	100
new group factors in	f_2	50	1 to 5	50	74	10	1	0
each set of 5 items:	.	.	0 to 5			.	.	.
$\frac{n}{5}+1 \leq k \leq \frac{n+4}{5}+1$	.	.	"			.	.	.
	.	.	"			.	.	.
	f_k	50 (if present)	"			10	1	0
	$\Sigma f_2 \cdots f_k$			50	74	48	21	0
	f_{U_i}	41	1					
	Σf_{U_i}		n	41	12	8	3	0
				$n=1$	$n=6$	$n=26$	$n=100$	$n\to\infty$
One general factor,	f_1	9	n	9	22	25	26	26
two group factors	f_2	50 or 0	$n/2$	50	31	35	36	37
each in half the	f_3	0 or 50	$n/2$	0	31	35	36	37
items	f_{U_i}	41	1					
	Σf_{U_i}		n	41	17	4	1	0

these also account for an increasing and perhaps substantial portion of the variance. But when each factor other than the first is present in only a few items, the general factor accounts for the lion's share of the variance as the test reaches normal length. We shall return to the implications of this for test design and for homogeneity theory.

Next, however, we apply this to coefficients of equivalence. We may study the composition of half-tests just as we have studied the total test. And we may also examine the composition of C_{ab}, the between-halves covariance. In Table 6, we consider first the test where there is a general factor and two group factors. If the test is divided into halves such that every item is factorially identical to its opposite number, save for the unique factor in each, the covariance C_{ab} nonetheless depends primarily upon the general-factor terms. Note, for example, the twenty-item test. Two-thirds of the covariance terms are the result of item similarity in the general factor. Suppose that these general factor terms are about equal in size. Then, should the test be split differently, the covariance would be reduced to the extent that more than half the items loaded with (say) factor 2 fall in the same half, but even the most drastic possible departure from the parallel split would reduce the covariance by only one-third of its terms. In the event that the group-factor loadings in the items are larger than the general-factor loadings, the size of the covariance is reduced by more than one-third. It is in this case that the parallel split has special advantage: where a few group factors are present and have loadings in the items larger than the general factor does.

The nature of the split has even less importance for the pattern where each factor is found in but a few items. Suppose, for example, that we are dealing with the 60-item test containing 15 factors in four items each. Then suppose that it is so very "badly" split that items containing 5 of the factors were assigned only to one of the half-tests, and items containing the second 5 factors were assigned to the other half-test. This would knock out 40 terms from the between-halves covariance, but such a shift would reduce the covariance only by 40/960 of its terms. Only in the exceptional conditions where general factor loadings are miniscule or where they vary substantially would different splits of such a test produce marked differences in the covariance.

It follows from this analysis that marked variation in the coefficients obtained when a test is split in several ways can result only when

TABLE 6
Composition of the Between-Halves Covariance for Tests of Certain Patterns

Pattern	Common Factors	No. of Items Having Non-Zero Loadings in Factor	No. of Terms Representing Each Factor in Between-Halves Covariance (ΣC_{ij}) When an Ideal Split is Made, for Varying Numbers of Items			
			$n=2$	$n=8$	$n=20$	$n=60$
One general factor,						
two group factors	1	n	1	16	100	900
each in half the	2	$n/2$		4	25	225
items	3	$n/2$		4	25	225
Total No. of Terms in C_{ab}			1	24	150	1350
Total No. of Terms in V_t			8	104	620	5460
			$n=2$	$n=8$	$n=20$	$n=60$
One general factor,	1	n	1	16	100	900
new group factors in	2	4	1	4	4	4
each set of 5 items:	3	4		4	4	4
$\frac{n}{5}+1 \leq k \leq \frac{n+4}{5}+1$	$4 \cdots k$	4 each			4 each	4 each
Total No. of Terms in C_{ab}			2	24	120	960
Total No. of Terms in V_t			10	104	500	3900

(a) a few group factors have substantial loadings in a large fraction of the items or

(b) when first-factor loadings in the items tend to be very small or where they vary considerably. Even these conditions are likely to produce substantial variations only when the variance of a test is contributed to by only a few items.

In the experimental tests studied by Clark, by Cronbach, and in the present study, general-factor loadings were probably greater, on the whole, than group-factor loadings. Moreover, none of the tests seems to have been divisible into large blocks of items each representing one group factor. (Such large "lumps" of group factor content are most often found in tests broken into subtests, viz., the Number Series, Analogies, and other portions of the ACE Psychological examination.)

This establishes on theoretical grounds the fact that for certain common types of test, there is likely to be negligible variation among split-half coefficients. Therefore α, the mean coefficient, represents such tests as well as any parallel split.

This interpretation differs from the Wherry-Gaylord conclusion (38) that "the Kuder-Richardson formula tends to underestimate the true reliability by the ratio $(n - K)/(n - 1)$ when the number of factors, K, is greater than one." They arrive at this by highly restrictive assumptions: that all factors are present in an equal number of items, that no item contains more than one factor, that there is no general factor, and that all items measuring a factor have equal variances and covariances. This type of test would never be intended to yield a psychologically interpretable score. For psychological tests where the intention is that all items include the same factor, our development shows that the quoted statement does not apply.

The problem of differential weighting has been studied repeatedly, the clearest mathematical analyses being those of Richardson (30) and Burt (4). This problem is closely related to our own study of test composition. Making different splits of a test is essentially the same as weighting the component items differently. The conditions under which split-half coefficients differ considerably are identical to those where differential weighting of components alters a total score appreciably: few components, lack of general factor or variation in its loadings, large concentrations of variance in group factors. The more formal mathematical studies of weighting lead to the same conclusions as our study of special cases of test construction.

4. *How is α related to the homogeneity, internal consistency, or*

*saturation of a test?** During the last ten years, various writers (12, 19, 27) directed attention to a property they refer to as homogeneity, scalability, internal consistency, or the like. The concept has not been sharply defined, save in the formulas used to evaluate it. The general notion is clear: In a homogeneous test, the items measure the same things.

If a test has substantial internal consistency, it is psychologically interpretable. Two tests, composed of different items of this type, will ordinarily give essentially the same report. If, on the other hand, a test is composed of groups of items, each measuring a different factor, it is uncertain which factor to invoke to explain the meaning of a single score. *For a test to be interpretable, however, it is not essential that all items be factorially similar.* What is required is that a large proportion of the test variance be attributable to the principal factor running through the test (37).

α estimates the proportion of the test variance due to all common factors among the items. That is, it reports how much the test score depends upon general and group, rather than item specific, factors. If we assume that the mean variance in each item attributable to common factors ($\overline{\sum_z \sigma_i^2 f_{zi}^2}$) equals the mean interitem covariance $\overline{\sum_z (\sigma_i \sigma_j f_{zi} f_{zj})}$,

$$\frac{1}{n}\sum_z \sum_i \sigma_i^2 f_{zi}^2 = \frac{2}{n(n-1)} \sum_i \sum_j C_{ij} = \frac{2}{n(n-1)} C_t. \qquad (39)$$

$$\sum_z \sum_i \sigma_i^2 f_{zi}^2 = \frac{2}{n-1} C_t, \qquad (40)$$

and the total variance (item variance plus covariance) due to common factors is $2\frac{n}{n-1}C_t$. Therefore, from (14), α is the proportion of test variance due to common factors. Our assumption does not hold true when the interitem correlation matrix has rank higher than one. Normally, therefore, α underestimates the common-factor variance, but not seriously unless the test contains distinct clusters.

The proportion of the test variance due to the first factor among the items is the essential determiner of the interpretability of the

*Several of the comments made in the following sections, particularly regarding Loevinger's concepts, were developed during the 1949 APA meetings in a paper by Humphreys (21) and in a symposium on homogeneity and reliability. The thinking has been aided by subsequent discussions with Dr. Loevinger.

scores. α is an upper bound for this. For those test patterns described in the last section, where the first factor accounts for the preponderance of the common-factor variance, α is a close estimate of first-factor concentration.

α applied to batteries of tests or subtests. Instead of regarding α as an index of *item* consistency, we may apply it to questions of *subtest* consistency. If each subtest is regarded as an "item" composing the test, formula (2) becomes

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum V_{\text{subtests}}}{V_{\text{test}}}\right). \tag{41}$$

Here n is the number of subtests. If this formula is applied to a test or battery composed of separate subtests, it yields useful information about the interpretability of the composite. Under the assumption that the variance due to common factors within each subtest is on the average equal to the mean covariance between subtests, α indicates what proportion of the variance of the composite is due to common factors among the subtests. In many instruments the subtests are positively correlated and intended to measure a general factor. If the matrix of intercorrelations is approximately hierarchical, so that group factors among subtests are small in influence, α is a measure of first-factor concentration in the composite.

Sometimes the variance of the test is not immediately known, but correlations between subtests are known. In this case one can compute covariances ($C_{ab} = \sigma_a \sigma_b r_{ab}$), or the variance of the composite (V_t is the sum of subtest variances and covariances), and apply formula (41). But if subtest variances are not at hand, an inference can be made directly from correlations. If all subtests are assigned weights such that their variances are equal, i.e., they make equal contributions to the total,

$$\alpha = \frac{n}{n-1}\left(\frac{2\sum\limits_{i}\sum\limits_{j} r_{ij}}{n + 2\sum\limits_{i}\sum\limits_{j} r_{ij}}\right); \; (i = 1, 2, \cdots n-1; j = i+1, \cdots n). \tag{42}$$

Here i and j are subtests, of which there are n. This formula tells what part of the total variance is due to the first factor among the subtests, when the weighted subtest variances are equal.

A few applications will suggest the usefulness of this analysis. The California Test of Mental Maturity, Primary, has two part scores, Language and Non-Language. For a group of 725, according to the test authors, these scores correlate .668. Then, by (42), α, the common-factor concentration, is .80. Turning to the Primary Mental Abilities Tests, we have a set of moderate positive correlations reported when these were given to a group of eighth-graders (35). The question may be asked: How much would a composite score on these tests reflect common elements rather than a hodgepodge of elements each specific to one subtest? The intercorrelations suggest that there is one general factor among the tests. Computing α on the assumption of equal subtest variances, we get .77. The total score is loaded to this extent with a general intellective factor. Our third illustration relates to four Air Force scores related to carefulness. Each score is the count of number *wrong* on a plotting test. The four scores have rather small intercorrelations (15, p. 687), and each score has such low reliability that its use alone as a measure of carefulness is not advisable. The question therefore arises whether the tests are enough intercorrelated that the general factor would cumulate in a preponderant way in their total. The sum of the six intercorrelations is 1.76. Therefore α is .62. I.e., 62% of the variance in the equally weighted composite is due to the common factor among the tests.

From this approach comes a suggestion for obtaining a superior coefficient of equivalence for the "lumpy" test. It was shown that a test containing distinct clusters of items might have a parallel-split coefficient appreciably higher than α. If so, we should divide the test into subtests, each containing what appears to be a homogeneous group of items. α is computed for each subtest separately by (2). Then $\sigma_i^2\alpha$ gives the covariance of each cluster with the opposite cluster in a parallel form, and the covariance between subtests is an estimate of the covariance of similar pairs "between forms." Hence

$$r_{t_1 t_2} = \frac{\sum_i \sum_j \sigma_i \sigma_j r_{ij}}{V_t}; \quad (i=1,2,\cdots n; j=1,2,\cdots n), \qquad (43)$$

where α_i is entered for r_{ii}, i and j being subtests. To the extent that $\bar{\alpha}_i$ is higher than the mean correlation between subtests, the parallel-forms coefficient will be higher than α_t computed from (2).

The relationships developed are summarized in Figure 3. α falls somewhere between the proportion of variance due to the first factor and the proportion due to all common factors. The blocks representing "other common factors" and "item specifics" are small, for tests not containing clusters of items with distinctive content.

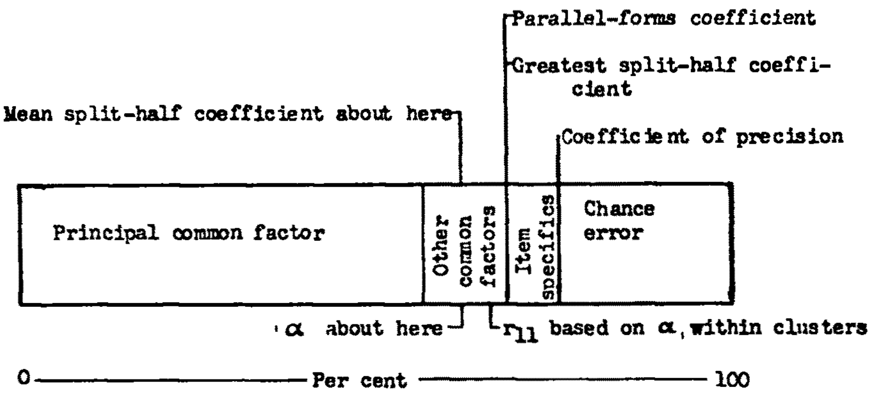
Parallel-forms coefficient
Greatest split-half coefficient
Mean split-half coefficient about here
Coefficient of precision
Principal common factor
Other common factors
Item specifics
Chance error
α about here
r11 based on α, within clusters
0
Per cent
100

FIGURE 3
Certain Coefficients related to the Composition of the Test Variance.

An index unrelated to test length. Conceptually, it seems as if the "homogeneity" or "internal consistency" of a test should be independent of its length. A gallon of homogenized milk is no more homogeneous than a quart. α increases as the test is lengthened, and so to some extent do the Loevinger-Ferguson homogeneity indices. We propose to obtain an indication of interitem consistency by applying the Spearman-Brown formula to α_t, thereby estimating the mean correlation between items. The formula is entered with the reciprocal of the number of items as the multiple of test length. The formula can be simplified to

$$\bar{r}_{ij(est)} = \frac{\alpha}{n + (1 - n)\alpha} \qquad (44)$$

or (cf. 24, p. 213 and 30, p. 387),

$$\bar{r}_{ij(est)} = \frac{1}{n - 1} \cdot \frac{V_t - \sum V_i}{\sum V_i}. \qquad (45)$$

$\bar{r}_{ij(est)}$ (r bar) is the correlation required, among items having equal variances and equal covariances, to obtain a test of length n having common-factor concentration α. $\bar{r}_{ij(est)}$ or its special case $\bar{\phi}$ for dichotomously-scored items is recommended as an overall index of internal consistency, if one is needed. It is independent of test length. It is not, in my opinion, important for a test to have a high $\bar{r}$ if α is high. Woodbury's "standard length" (39) is an index of internal consistency which can be derived from $\bar{r}_{ij}$ and has the same advantages and limitations. n_i, the standard length, is the number of items which yields an α of .50 . Then

$$n_i = \frac{1 - \bar{r}_{ij}}{\bar{r}_{ij}}. \tag{46}$$

If $\bar{r}$ is high, α is high. But α may be high even when items have small intercorrelations. If $\bar{r}$ is low, the test may be a smooth mixture of items all having low intercorrelations. In this case, each item would have some loading with the general factor and if the test is long α could be high. Such items are illustrated by very difficult psychophysical discriminations such as a series of near-threshold speech signals to be interpreted; with enough of these items we have a highly satisfactory measuring instrument. In fact, save for random error of performance, it may be unidimensional. A low value of $\bar{r}$ may instead indicate a lumpy test composed of discrete and homogeneous subtests. Guttman (34, p. 176n.) describes a questionnaire of this type. The concept of homogeneity has no particular meaning for a "lumpy" test. It is logically meaningless to inquire whether a set of ten measures of physical size plus ten intercorrelated vocabulary items is more homogeneous than twenty slightly correlated biographical questions. A high $\bar{r}$ is sufficient but not necessary evidence that the test lacks important group factors. When $\bar{r}$ is low, only a study of correlations among items or trial clusters of items shows whether the test can be broken into more homogeneous subtests.

Comparison with the index of reproducibility. Guttman's coefficient of reproducibility has appeared to some reviewers (Loevinger, 28; Festinger, 13) as an *ad hoc* index with no mathematical rationale. It may therefore be worthwhile to note that this coefficient can be approximated by a mathematical form which makes clear what it measures. The correlation of any two-choice item with a total score on a test may be expressed as a phi coefficient, and this is common in conventional item analysis. Guttman dichotomizes the test scores at a cutting point selected by inspection of the data. We will get similar results if we dichotomize scores at that point which cuts off the same proportion of cases as pass the item under study. (Our ϕ_{it} will be less in some cases than it would be if determined by Guttman's inspection procedure.) Simple substitution in Guttman's definition (34, p. 117) leads to

$$\underline{R} \doteq \overline{1 - 2\sigma_i{}^2(1 - \phi_{it})}, \tag{47}$$

where the approximation is introduced by the difference in ways of

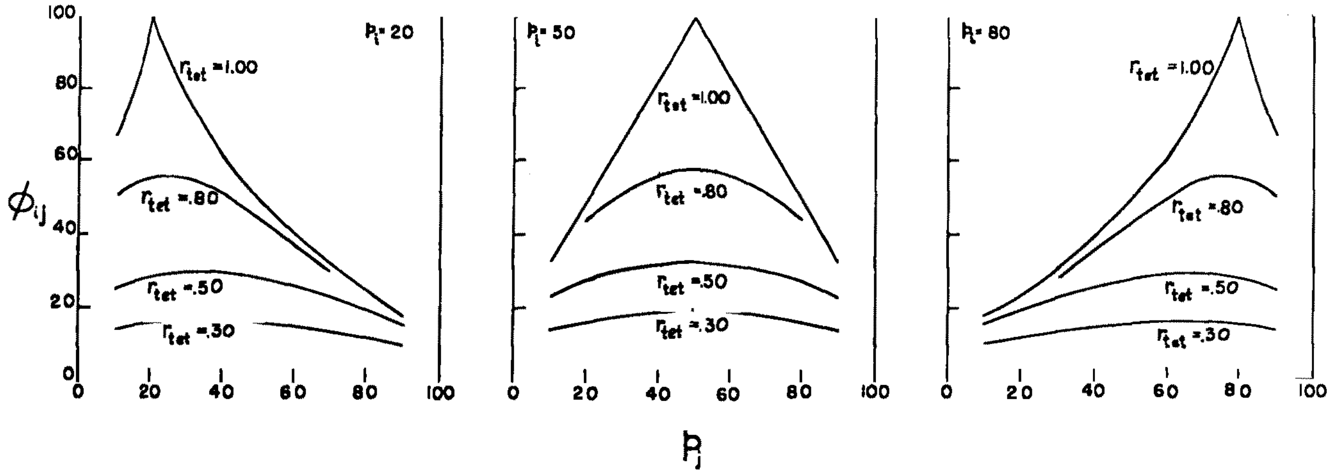

FIGURE 4
Relation of ϕ_{ij} to p_i and p_j for Several Levels of Correlation.

dichotomizing. The actual R obtained by Guttman will be larger than that from (47). For multiple-alternative items, a similar but more complex formula involving the phi coefficient of the alternative with the test is required to approximate Guttman's result. R is independent of test length; if a Guttman scale is divided into equivalent portions, the two halves will have the same R as the original test. In this respect, R is most comparable to our $\bar{r}$. Both ϕ_{it} and $\bar{r}$ are low, so long as items are unreliable or contain substantial specific factors.

5. *Is the usefulness of α limited by properties of the phi coefficient between items having unequal difficulties?* The criticism has been made, most vehemently by Loevinger (27), that α is a poor index because, being based on product-moment correlations, it cannot attain unity unless all items have distributions of the same shape. For the pass/fail item, this requires that all p_i be equal. The inference is drawn that since the coefficient cannot reach unity for such items, α and $\bar{r}$ do not properly represent homogeneity.

There are two ways of examining this criticism. The simpler is empirical. The alleged limitation upon the product-moment coefficient has no practical effect upon the coefficient, for items of the sort customarily employed in psychological tests. To demonstrate this, we consider the change in ϕ with changes in item difficulty. To hold constant the relation between the "underlying traits," we fix the tetrachoric correlation. When the tetrachoric coefficient is .30, $p_i = .50$ and p_j ranges from .10 to .90, ϕ_{ij} ranges only from .14 to .19. Figure 4 shows the relation of ϕ_{ij} to p_i and p_j for three levels of correlation: $r_{tet} = .30$, $r_{tet} = .50$, and $r_{tet} = .80$. The correlation among items in psychometric tests is ordinarily below .30. For example, even for a five-grade range of talent, the $\bar{\phi}_{ij}$ for the California Test of Mental Maturity subtests range only from .13 to .25. That is, for tests having the degree of item intercorrelation found in present practice, ϕ is very nearly constant over a wide range of item difficulties.

TABLE 7

Variation in Certain Indices of Interitem Consistency with Changes in Item Difficulty (Tetrachoric Correlation Held Constant)

p_i	.50	.50	.50	.50	.50	.50	.50	.50	.50
p_j	→.00	.10	.20	.40	.50	.60	.80	.90	→1.00
$r_{ij\,tet}$	.30	.30	.30	.30	.30	.30	.30	.30	.30
ϕ_{ij}	→.00	.14	.17	.19	.19	.19	.17	.14	→.00
H_{ij}	→1.00	.42	.34	.23	.19	.23	.34	.42	→1.00

Examining Loevinger's proposed coefficient of homogeneity (29),

$$H_{ij} = \phi_{ij}/\phi_{ij(\max)} , \tag{48}$$

we find that *it* is markedly affected by variations in item difficulty. One example is worked out in Table 7. As many investigators including Loevinger have noted, Guttman's R is drastically affected by item difficulty. For any single item, R must be greater than p_i or q_i, whichever is greater. Evidently the indices of homogeneity which might replace $\bar{\phi}$ suffer more from the effects of differences in difficulty than does the phi coefficient.

Further evidence on the alleged limitation of α is obtained by preparing four hypothetical 45-item tests. In each case, all $r_{ij(\text{tet})}$ are fixed at .30. Phi coefficients reflect both heterogeneity in content and heterogeneity in difficulty. To assess the effect of the latter heterogeneity upon $\bar{\phi}$ and α, we compared one test of uniform item difficulty, where all heterogeneity is in content, with another where "heterogeneity due to difficulty" was allowed to enter. As Table 8 indicates, even when extreme ranges of item difficulty are allowed, neither $\bar{\phi}$ nor α is affected in any practically important way. For tests where item difficulties are higher, or correlations are lower, the effect would be even more negligible.

TABLE 8

Comparison of $\bar{\phi}$ and α for Hypothetical 45-Item Tests With and Without "Heterogeneity Due to Item Difficulty"

Test	Distribution of Difficulties	Range of p_i	$\bar{p}_i$	$\bar{\phi}$	Diff.	α	Diff.
A	Normal	.20 to .80	.50	.181	*.011*	.909	*.005*
A′	Peaked	.50	.50	.192		.914	
B	Normal	.10 to .90	.50	.176	*.016*	.906	*.008*
B′	Peaked	.50	.50	.192		.914	
C	Normal	.50 to .90	.70	.170	*.011*	.902	*.007*
C′	Peaked	.70	.70	.181		.909	
D	Rectangular	.10 to .90	.50	.153	*.039*	.892	*.022*
D′	Peaked	.50	.50	.192		.914	

Still another small study leading to the same essential conclusion was made by examining a "perfect scale," where all p_{ij} equal $\phi_{ij(\max)}$. Items were placed at five difficulty levels, the p_i being .50, .58, .71, .80, and .89. Then the correlations (phis) of items range from 1.00 (at same level) to .85 (highest between levels) to .36. In a test of only five items, α reaches .86 . This is the maximum α could have, for this set of 5 items and specified p_i . As the number of items increases, α rises toward 1.00 . Thus, for 10 items, two at each level $\alpha_{\max} = .951$; for 20 items, .977 . It follows that even if items are much more homogeneous in content than present tests and much freer from error, the cumulative properties of covariance terms make the failure of all ϕ's to reach unity of next-to-no importance. $\alpha_{\max}$ would be lower if difficulties range over the full scale, but the same principle holds. α is a good measure of common-factor concentration, for tests of reasonable length, in spite of the fact that it falls short of 1.00 if items vary in difficulty.

In the case of the perfect scale, of course, $\bar{\phi}$ does fall well short of unity and for such tests it does not reflect the homogeneity in content. From the five-item case just considered, $\bar{\phi}$ is .54 .

The second way to analyze this criticism is to examine the nature of redundancy (using a term from Shannon's information theory, 32). If two items repeat the same information, they are totally redundant. Thus, if one item divides people 50/50, and the second item does also, the two items always placing exactly the same people together, the second item gives no new information about individual differences. (Cf. Tucker, 36). Suppose, though, that the second item is passed by 60 per cent of the subjects. Even if $r_{ij(\text{tet})} = 1.00$, this second item conveys new information because it discriminates among the fifty people who failed the first item. A five-item test where all items have perfect tetrachoric intercorrelations, and the p_i are .40 , .45 , .50 , .55 , .60 , is perfectly homogeneous (a la Guttman, Loevinger, et al). So is a ten-item test composed of these items plus five others whose p's are .30 , .35 , .65 , .70 , .75 . The two tests are not equivalent in measuring power, however; the second makes a much greater number of discriminations. Because there is less redundancy, the longer test has a lower $\bar{\phi}$.

From the viewpoint of information theory, we should be equally concerned with heterogeneity in content and heterogeneity in difficulty. We get one bit of information when we place the person as above the mean in (say) pitch discrimination. Now with another item or set of items, we might place him relative to the mean in visual acuity. The two tests together place him in one of four categories. If our second test had been a further measure of pitch, placing the subject above or below the 75th percentile, then the two tests would have placed him in one of four categories. Either set of tests gives the same amount of information. Which information we most want depends on practical considerations.

The phi coefficient reports whether a second item gives new information that the first does not. Then a tetrachoric r must be computed to determine if the new information relates to a new content dimension or to a finer discrimination on the same content dimension. If the phi coefficient between true scores is 1.00, redundancy is complete and there is no new information. Redundancy is desirable when accuracy of a single item is low. To test whether men can hear a 10-cycle difference, the best way is to use a large number of items of just that difficulty. Such items usually also discriminate to some degree at other points on the scale, but cannot give information about ability at the 5-cycle level if a single item is extremely reliable. With very accurate items a pitch test which is not homogeneous will be better for differentiation all along the scale. The "factors" found by Ferguson (11) due to the higher correlation (redundancy) of items with equal difficulty need not be regarded as artifacts (38).* These "difficulty factors" are factors on which the test gives information and on which the tester may well want information. They are not "content factors," but they must be considered in test analysis. For example, if one regards pitch tests in this light, it is seen that a test containing 5-cycle items, 10-cycle items, and 15-cycle items will be slightly influenced by undesired factors, when the criterion requires discrimination only at the 15-cycle level. (Problems of this type occur in validating tests for selecting military personnel using detection apparatus). One would maximize the loading in the test of the group factor among 15-cycle items, to maximize validity. This factor is of course a mathematical factor, and not a property of the auditory machinery. While the mathematics is not clear, it seems very likely that the group factors found among phi coefficients are interchangeable with Guttman's "components of scale analysis" to which he gives serious psychological interpretation.

From this point of view, the phi coefficient which tells when items do and do not duplicate each other is a better index *just because* it does not reach unity for items of unequal difficulty. Phi and r_{tet} are both useful in test analysis. Brogden (1, pp. 199, 201) makes a similar point, although approaching the problem from another tack.

*It is not *necessary*, as Ferguson seems to think, for difficulty factors to emerge if product-moment correlations are used with multi-category variates. On *a priori* grounds, difficulty factors will appear only if the shapes of the distributions of the variates are different. In Ferguson's data it appears likely that the hardest and easiest tests were skewed in opposite directions.

Implications for Test Design

In view of the relations detailed above, we find it unnecessary to create homogeneous scales such as Guttman, Loevinger, and others have urged.

It is true that a test where all items represent the same content factor with no error of measurement is maximally interpretable. Everyone attaining the same score would mark items in the same way. Yet the question we really wish to ask is whether the individual differences in test score are attributable to the first factor within the test. If a large proportion of the score variance relates to this factor, the residue due to specific characteristics of the items little handicaps interpretability. It has been shown that a high first-factor saturation indicated by a high α can be attained by cumulating many items which have low correlations. The standard proposed by Ferguson, Loevinger, and Guttman is unreasonably severe, since it would rule out tests which do have high first-factor concentrations.

These writers seem to wish to infer the person's score on each item from his total score. This appears unimportant, but even if it were important, the interest would attach to predicting his *true* standing on the item, not his fallible obtained score. For the unreliable items used in psychological and educational tests, the aim of Guttman et al. will not be approached in practice. Perhaps sociological data have such greater reliability that prediction of obtained scores is tantamount to predicting true scores.

Increasing interpretability by lengthening a test is not without its disadvantages. Using more and more time to get at the same information employs the principle of redundancy (32). When a message is repeated over and over, it is easier to infer the true message even when there is substantial interference (item unreliability). But the more you repeat messages already transmitted, the less time is allowed for conveying other information. A set of redundant items can carry much less information than a set of independent items. In other words, when we lengthen certain tests or subtests to make their scores more interpretable, we sacrifice the possibility of obtaining separate measures of additional factors in the same time.

From the viewpoint of both interpretability and efficient prediction of criteria, the smallest element on which a score is obtained should be a set of items having a substantial α and not capable of division into discrete item clusters which themselves have high α. Such separately interpretable tests can sometimes be combined into an interpretable composite, as in the case of the PMA tests. Although it is believed that the test designer should seek interitem consistency, and judge the effectiveness of his efforts by the coefficient α, the pure scale should not be viewed as an ideal. It should be remembered that Tucker (36) and Brogden (1) have demonstrated that increases in internal consistency may lead to decreases in the product-moment validity coefficient when the shape of the test-score distribution differs from that of the criterion distribution.

Summary

1. Formulas for split-half coefficients of equivalence are compared, and those of Rulon and Guttman are advocated for practical use rather than the Spearman-Brown formula.

2. α, the general formula of which Kuder-Richardson formula 20 is a special case, is found to have the following important meanings:

(a) α is the mean of all possible split-half coefficients.

(b) α is the value expected when two *random* samples of items from a pool like those in the given test are correlated.

(c) α is a lower bound for the coefficient of precision (the instantaneous accuracy of this test with these particular items). α is also a lower bound for coefficients of equivalence obtained by simultaneous administration of two tests having matched items. But for reasonably long tests not divisible into a few factorially-distinct subtests, α is nearly equal to "parallel-split" and "parallel-forms" coefficients of equivalence.*

(d) α estimates, and is a lower bound to, the proportion of test variance attributable to common factors among the items. That is, it is an index of common-factor concentration. This index serves purposes claimed for indices of homogeneity. α may be applied by a modified technique to determine the common-factor concentration among a battery of subtests.

*W. G. Madow suggests that the amount of disagreement between two random or two planned samples of items from a larger population of items could be anticipated from sampling theory. The person's score on a test is a sample mean, intended to estimate the population mean or "true score" over all items. The variance of such a mean from one sample to another decreases rapidly as the sample is enlarged by lengthening the test, whether samples are drawn at random or are drawn after stratifying the universe as to difficulty and content. The conditions under which the radom splits correlate about as highly as parallel splits are those in which stratified sampling has comparatively little advantage. Madows comment has implications also for the preparation of comparable forms of tests and for developing objective methods of selecting a sample of items to represent a larger set of items so that the variance of the difference between the score based on the sample and the score based on the universe of items is as small as possible.

(e) α is an upper bound to the concentration in the test of the first factor among the items. For reasonably long tests not divisible into a few factorially-distinct subtests, α is very little greater than the exact proportion of variance due to the first factor.

3. Parallel-splits yield coefficients little larger than random splits, unless tests contain large blocks of items representing group factors. For such tests, α computed for separate blocks and combined by a special formula gives a satisfactory estimate of first-factor concentration.

4. Interpretability of a test score is enhanced if the score has a high first-factor concentration. A high α is therefore to be desired, but a test need not approach a perfect scale to be interpretable. Items with quite low intercorrelations can yield an interpretable scale.

5. A coefficient $\overline{r}_{ij}$ (or $\overline{\phi}_{ij}$) is derived which is the intercorrelation required, among items with equal intercorrelations and variances, to reproduce a test of n items having common-factor concentration α. $\overline{\phi}$, as a measure of item interdependence, draws attention to heterogeneity in both difficulty and content factors. Heterogeneity in test difficulty merits the attention of the test designer, since the validity of the test may be increased by capitalizing on "difficulty factors" present in the criterion.

6. To obtain subtest scores for interpretation or to be weighted in an empirical composite, the ideal set of items is one having a substantial α and not further divisible into a few discrete smaller blocks of items.

REFERENCES

1. Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 1946, **11**, 197-214.
2. Brown, W. Some experimental results in the correlation of mental abilities. *Brit. J. Psychol.*, 1910, **3**, 296-322.
3. Brownell, W. A. On the accuracy with which reliability may be measured by correlating test halves. *J. exper. Educ.*, 1933, **1**, 204-215.
4. Burt, C. The influence of differential weighting. *Brit. J. Psychol., Stat. Sect.*, 1950, **3**, 105-128.
5. Clark, E. L. Methods of splitting vs. samples as sources of instability in test-reliability coefficients. *Harvard educ. Rev.*, 1949, **19**, 178-182.
6. Coombs, C. H. The concepts of reliability and homogeneity. *Educ. psychol. Meas.*, 1950, **10**, 43-56.
7. Cronbach, L. J. On estimates of test reliability. *J. educ. Psychol.*, 1943, **34**, 485-494.

8. Cronbach, L. J. A case study of the split-half reliability coefficient. *J. educ. Psychol.*, 1946, 37, 473-480.

9. Cronbach, L. J. Test "reliability": its meaning and determination. *Psychometrika*, 1947, 12, 1-16.

10. Dressel, P. L. Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika*, 1940, 5, 305-310.

11. Ferguson, G. The factorial interpretation of test difficulty. *Psychometrika*, 1941, 6, 323-329.

12. Ferguson, G. The reliability of mental tests. London: Univ. of London Press, 1941.

13. Festinger, L. The treatment of qualitative data by "scale analysis." *Psychol. Bull.*, 1947, 44, 149-161.

14. Goodenough, F. L. A critical note on the use of the term "reliability" in mental measurement. *J. educ. Psychol.*, 1936, 27, 173-178.

15. Guilford, J. P., ed. Printed classification tests. Report No. 5, Army Air Forces Aviation Psychology Program. Washington: U. S. Govt. Print. Off., 1947.

16. Guilford, J. P. Fundamental statistics in psychology and education. Second ed. New York: McGraw-Hill, 1950.

17. Guilford, J. P., and Michael, W. B. Changes in common-factor loadings as tests are altered homogeneously in length. *Psychometrika*, 1950, 15, 237-249.

18. Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

19. Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, 10, 255-282.

20. Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, 6, 153-160.

21. Humphreys, L. G. Test homogeneity and its measurement. *Amer. Psychologist*, 1949, 4, 245.

22. Jackson, R. W., and Ferguson, G. A. Studies on the reliability of tests. Bull. No. 12, Dept. of Educ. Res., University of Toronto, 1941.

23. Kelley, T. L. Note on the reliability of a test: a reply to Dr. Crum's criticism. *J. educ. Psychol.*, 1924, 15, 193-204.

24. Kelley, T. L. Statistical method. New York: Macmillan, 1924.

25. Kelley, T. L. The reliability coefficient. *Psychometrika*, 1942, 7, 75-83.

26. Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.

27. Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, 61, No. 4.

28. Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychol. Bull.*, 1948, 45, 507-529.

29. Mosier, C. I. A short cut in the estimation of split-halves coefficients. *Educ. psychol. Meas.*, 1941, 1; 407-408.

30. Richardson, M. Combination of measures, pp. 379-401 in Horst, P. (Ed.) The prediction of personal adjustment. New York: Social Science Res. Council, 1941.

31. Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. *Harvard educ. Rev.*, 1939, 9, 99-103.

32. Shannon, C. E. The mathematical theory of communication. Urbana: Univ. of Ill. Press, 1949.

33. Spearman, C. Correlation calculated with faulty data. *Brit. J. Psychol.*, 1910, 3, 271-295.
34. Stouffer, S. A., et. al. Measurement and prediction. Princeton: Princeton Univ. Press, 1950.
35. Thurstone, L. L., and Thurstone, T. G. Factorial studies of intelligence, p. 37. Chicago: Univ. of Chicago Press, 1941.
36. Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika*, 1946, 11, 1-13.
37. Vernon, P. E. An application of factorial analysis to the study of test items. *Brit. J. Psychol., Stat. Sec.*, 1950, 3, 1-15.
38. Wherry, R. J., and Gaylord, R. H. The concept of test and item reliability in relation to factor pattern. *Psychometrika*, 1943, 8, 247-264.
39. Woodbury, M. A. On the standard length of a test. Res. Bull. 50-53, Educ. Test. Service, 1950.

Manuscript received 12/28/50
Revised manuscript received 2/28/51