PSYCHOMETRIKA—VOL. 10, NO. 4
DECEMBER, 1945

A BASIS FOR ANALYZING TEST-RETEST RELIABILITY*

LOUIS GUTTMAN
DEPARTMENT OF SOCIOLOGY AND ANTHROPOLOGY
CORNELL UNIVERSITY

Three sources of variation in experimental results for a test are distinguished: trials, persons, and items. Unreliability is defined only in terms of variation over trials. This definition leads to a more complete analysis than does the conventional one; Spearman's contention is verified that the conventional approach—which was formulated by Yule—introduces unnecessary hypotheses. It is emphasized that at least two trials are necessary to estimate the reliability coefficient. This paper is devoted largely to developing *lower bounds* to the reliability coefficient that can be computed from but a *single trial*; these avoid the experimental difficulties of making two independent trials. Six different lower bounds are established, appropriate for different situations. Some of the bounds are easier to compute than are conventional formulas, and all the bounds assume less than do conventional formulas. The terminology used is that of psychological and sociological testing, but the discussion actually provides a general analysis of the reliability of the sum of n variables.

CONTENTS

PART I *The Definition of Reliability*

PART II *Lower Bounds for the Reliability Coefficient*

*The writer is indebted to the members of his statistical seminar, to Professor Mark Kac, and to Professor Samuel A. Stouffer and his staff in the Research Branch, Information and Education Division, War Department, for their helpful comments on this paper.

PART I. *The Definition of Reliability*

1. *Introduction.* It is now over forty years since Spearman first wrote about errors of observation (6). These errors, as he pointed out, have the remarkable property that they attenuate a correlation coefficient in a manner that cannot be remedied by increasing the number of individuals upon whom the correlation is based. Spearman's work has had great influence on much research in the psychological and social sciences. The reliability of instruments like achievement tests and attitude questionnaires has become a standard problem for investigation.

The analytical formulation of the problem that is conventional today* is that originally submitted by Yule in a letter to Spearman (7). It assumes that an observation consists of a "true score" plus an error, where the error is assumed to have a zero mean over individuals and is assumed to correlate zero with the "true score" over individuals. Further assumptions concern zero covariances between errors, the covariances being taken over individuals. In publishing this formulation, which has become classical, Spearman remarks that he believes Yule makes more assumptions than are needed.

The present paper is devoted to a reformulation of the analysis of reliability according to what seems to have been Spearman's original purpose. It bears out Spearman's contention that the conventional formulation, while simple in its algebra, encumbers the analysis with unnecessary hypotheses. Also, the present paper produces new formulas that not only assume less than do the conventional ones, but are simpler to compute in practice.

The problem of reliability is of course not peculiar to psychology or sociology, but pervades all the sciences. In dealing with empirical data in any field, the question should be raised: if the experiment were to be repeated, how much variation would there be in the results? One of Spearman's important contributions has been to focus attention on the reliability of the *sum* of a number of variables. This is especially appropriate for psychological and sociological instruments like achievement and attitude tests which are scored by adding up item values. Our treatment stresses the reliability of a sum.

* See, for example, (1), p. 411.

The formulation to be presented here differs from the conventional one in the following respects:

(1) Error of observation is defined explicitly for each person on each item for each trial in a universe of trials. Thus the three sources of variation in an experiment are kept distinct: trials, persons, and items. Unreliability is defined as variation over trials.

(2) Using this definition, no assumptions of zero means for errors or zero correlations are needed to prove that the total variance of the test is the sum of the error variance and the variance of expected scores; this relationship between variances is an algebraic identity. Therefore, the reliability coefficient is defined without assumptions of independence as the complement of the ratio of error variance to total variance.

(3) A major emphasis of this paper is that the reliability coefficient cannot in general be estimated from but a single trial—that items do not replace trials. If two trials are experimentally independent, then we show that the correlation between two trials is, with probability of unity, equal to the reliability coefficient.

(4) As is well known, there may be great practical difficulties in making two independent trials; therefore our principal focus is on *what information can be obtained from a single trial.* We find that *lower bounds* to the reliability coefficient can be computed from a single trial. Six different lower bounds are derived, appropriate for different situations. Several of these bounds are as easy as or easier to compute than are conventional formulas, and all of the bounds assume less than do conventional formulas.

(5) To prove that bounds can be computed from a single trial, we use essentially one basic assumption: that the errors of observation are independent between items and between persons over the *universe of trials.* In the conventional approach, independence is taken over *persons* rather than trials, and the problem of observability from a single trial is not explicitly analyzed.

(6) We make no assumptions about the relationships between the items themselves, that is, as to what the relationships would be if there were no experimental error.

(7) Proof that bounds can be computed from but a single trial (and that the reliability coefficient itself can be computed from two independent trials) turns out to involve the notion of *convergence in the mean,* so that the results hold with *probability of unity.* This kind of analysis is required by the problem of reliability. The algebra in the last part of this paper may be somewhat tedious, but this should not obscure the easy formulas that emerge for use in practice.

The major practical results of this paper are the formulas for the lower bounds to the reliability coefficient, which can be computed from but a single trial. These bounds are listed in the next section, so that the reader who desires only working formulas can find them immediately. The formulas are for large samples; caution should be exercised in applying them to small samples.

The succeeding sections of Part I are devoted to laying down basic definitions and notation for the mathematical analysis. In Part II, various lower bounds to the reliability coefficient are derived in terms of parameters defined over persons and trials. Part III proves that these bounds can be computed from but a single trial.

2. *Working Formulas for the Lower Bounds.* A fundamental fact concerning unreliability is that, in general, it *cannot be estimated from only a single trial.* Two or more trials are needed to prove the existence of variation in the score of a person on an item, and to estimate the extent of such variation if there is any.

The experimental difficulties in obtaining independent trials have led to many attempts to estimate the reliability of a test from only a single trial by bringing in various hypotheses. Such hypotheses usually do not afford a real solution, since ordinarily *they cannot be verified* without the aid of at least two independent trials, which is precisely what they are intended to avoid.

An important result of this paper is to show that from a single trial, while it is not possible to estimate the reliability of a test, it is possible to set *lower bounds* to the reliability coefficient. In practice, such lower bounds will often be usefully greater than zero. It is assumed only that the items are experimentally independent, that the population of individuals is indefinitely large, and that the universe of (hypothetical) trials is indefinitely large. Since the working formulas are for large samples, they should be used with caution for small samples.

The reliability coefficient for the test will be denoted by ρ_t^2. In Part II of this paper are developed six lower bounds to ρ_t^2 in terms of parameters defined over all persons and trials; these bounds are denoted there by λ with distinguishing subscripts. In Part III, it is proved that each λ can be computed from but a single trial with probability unity. The computation for a λ from a single trial we shall denote here by an L with the corresponding subscript. The *L's* are the working formulas.

For each of the *L*'s, it is true with probability unity that

$$\underline{L} \leqq \rho_t^2 \leqq 1 \, .$$

The computations for the *L*'s are as follows.

For a *given trial,* let $s_1^2, s_2^2, \cdots, s_n^2$ be the variances over persons of the n items in the test, and let s_t^2 be the variance over persons of the sum of the items. The simple lower bound λ_1 (see §12 below) can be computed from the formula

$$L_1 = 1 - \frac{\sum\limits_{j=1}^{n} s_j^2}{s_t^2}.$$

Since L_3 below is a better bound (if $L_1 > 0$) and easily computed, L_1 will ordinarily not be used by itself in practice; but it is helpful in computing and comparing the various lower bounds.

A definitely better lower bound than L_1 is L_2, which requires computing the sum of squares of the covariances between items for the given trial; this sum of squares is denoted by C_2. The bound λ_2 (§13 below) is computed from the single trial by

$$L_2 = L_1 + \frac{\sqrt{\dfrac{n}{n-1} C_2}}{s_t^2}.$$

C_2 is the sum of $n(n-1)$ terms; but the covariances are equal in pairs since the covariance of item j with item k is equal to the covariance of item k with item j; therefore C_2 is simply twice the sum of the squares of the $n(n-1)/2$ *different* covariances.

λ_3 is derived in §14 by weakening λ_2 in order to save the labor of computing covariances. It is better than λ_1, if the latter is positive, since it is computed from the formula

$$L_3 = \frac{n}{n-1} L_1.$$

This is so easy to compute that L_1 need never be used by itself as a lower bound.

The relationship between these first three bounds is expressed by the inequalities:

$$L_1 < L_3 \leqq L_2,$$

the first inequality assuming that $L_1 > 0$.

A conventional formula that attempts to estimate the reliability coefficient by using a series of assumptions is the well known "corrected split-half coefficient." Our lower bound λ_4 (§15) resembles this coefficient. Two things, however, must be remembered: (a) L_4 assumes nothing additional to the assumptions stated above in this section in order to be used from a single trial, and (b) L_4 *underesti-*

mates the reliability coefficient. Furthermore, L_4 is *easier to compute* than is the "corrected split-half coefficient," since no correlation coefficient is explicitly computed. The formula for L_4 requires that the test be scored as two halves. The respective variances of the two parts for the single trial are denoted by s_a^2 and s_b^2, and the formula is

$$L_4 = 2\left(1 - \frac{s_a^2 + s_b^2}{s_t^2}\right).$$

L_4 is a lower bound *no matter how the test is split*. It is desirable, of course, to try to split the test in such a manner as to maximize L_4; L_4 will tend to be larger for halves which correlate more highly with each other.

If $s_a^2 = s_b^2$ for a particular splitting of the test, then L_4 is numerically equal to the "corrected split-half coefficient" for that split, but it is still a *lower bound* to the reliability coefficient. This may help explain why, in the past, the "correction for attenuation" has sometimes yielded a "correlation" greater than unity; reliability has often been underestimated by the conventional formula, so that attenuation has been overcorrected. Many tests are more reliable than they have been considered to be; and many low correlations have not necessarily been due to unreliability.

L_3 and L_4 are both easy to compute and will ordinarily be the most convenient of the lower bounds to use in practice. For the case of dichotomous scoring, where items have only the values zero and one ("wrong" and "right" in an achievement test), then L_3 is the easier of the two to compute, since then

$$s_j^2 \equiv p_j(1 - p_j),$$

where p_j is the proportion of people having the value of unity on the jth item (the proportion getting the jth item "right"). However, often it will be easy to find a splitting of the test that will yield an L_4 that will be substantially higher than L_3, so that it will be definitely preferable to compute L_4 in such cases.

Two further lower bounds that will sometimes be of use are λ_5 and λ_6. For the single trial, let C_{2j} be the sum of the squares of the covariances of item j with the remaining n—1 items, and let $\overline{C}_2$ be the largest of the C_{2j}. Then λ_5 (§16) can be computed from

$$L_5 = L_1 + \frac{2\sqrt{\overline{C}_2}}{s_t^2}.$$

L_5 will be greater than L_2, and hence L_3, for a test in which one item has large covariances with the other items compared with the covariances among those items. Otherwise, L_5 is less than or equal to L_2.

The sixth of our lower bounds is based on multiple correlation. For the single trial, let e_j^2 be the variance of the errors of estimate of item j from its linear multiple regression on the remaining $n-1$ items. Then λ_6 (§17) can be computed from

$$L_6 = 1 - \frac{\sum_{j=1}^{n} e_j^2}{s_t^2}.$$

L_6 will tend to be larger than L_2, and hence L_3, when the items have relatively low zero-order intercorrelations but high multiple correlations. Otherwise, L_6 will tend to be less than or equal to L_2.

The probability is unity that the reliability coefficient ρ_t^2 is *not smaller than the largest of* L_1, L_2, L_3, L_4, L_5, and L_6. And no matter what these lower bounds may be, they cannot disprove the hypothesis that $\rho_t^2 = 1$. At least two trials are needed to disprove such a hypothesis. A single trial can set a minimum to the reliability coefficient, but not a maximum less than unity.

3. *The Definition of Error.* A proper analysis of reliability must begin with a precise definition of error. The errors with which we are concerned are defined by an indefinitely large *universe of trials.* They are defined separately for each individual in a *population* for each *item* (variable) being observed.

Consider a set of n items. Let x_{ijk} be the observation of the ith individual on the jth item in the kth trial. The expected (mean) value for this individual on the item over all trials will be denoted by

$$X_{ij} = \underset{k}{E} x_{ijk}, \qquad (1)$$

where E denotes mean value (mathematical expectation) over the indicated subscript. The variance of the ith individual on the jth item over all trials is

$$\sigma^2_{x_{ij}} = \underset{k}{E} (x_{ijk} - X_{ij})^2. \qquad (2)$$

This may be called the *error variance of the* i*th individual on the* j*th item.*

The *test score,* or *total score,* of the ith person on the kth trial will be denoted by t_{ik}, and by definition,

$$t_{ik} = \sum_{j=1}^{n} x_{ijk}. \qquad (3)$$

The *expected* test score over all trials of the ith person will be denoted by

$$T_i = \underset{k}{E} t_{ik}. \qquad (4)$$

Taking expectations over trials of both members of (3), and using (1) and (4), yields

$$T_i = \sum_{j=1}^{n} X_{ij}. \tag{5}$$

The *error variance on the test* for the *i*th person is defined as

$$\sigma^2_{t_i} = \underset{k}{E}(t_{ik} - T_i)^2. \tag{6}$$

4. *The Variation Between Individuals and the Total Variation.* Thus far we have defined the variation *within* a person over the universe of trials. We shall also need to consider the variation *between* persons, which is done in terms of their expected scores over trials. The mean over persons of the expected test scores is

$$\mu_T = \underset{i}{E} T_i, \tag{7}$$

and the variance over persons of the expected test scores is

$$\sigma_T^2 = \underset{i}{E}(T_i - \mu_T)^2. \tag{8}$$

Finally, we need to define the total variation over all persons and trials. The general mean of the test over all persons and trials, μ_t, is of course equal to μ_T because

$$\mu_t = \underset{i}{E}\underset{k}{E} t_{ik} = \underset{i}{E} T_i = \mu_T .$$

The *total variance of the test* over all trials and all people is

$$\sigma_t^2 = \underset{i}{E}\underset{k}{E}(t_{ik} - \mu_t)^2. \tag{9}$$

Now,

$$\begin{aligned} \underset{k}{E}(t_{ik} - \mu_t)^2 &= \underset{k}{E}[(t_{ik} - T_i) + (T_i - \mu_T)]^2 \\ &= \underset{k}{E}(t_{ik} - T_i)^2 + (T_i - \mu_T)^2. \end{aligned}$$

Substituting the last member into (9), and remembering (6) and (8), we get the basic formula:

$$\sigma_t^2 = \underset{i}{E}\sigma^2_{t_i} + \sigma_T^2. \tag{10}$$

5. *The Definition of the Reliability Coefficient.* We shall define the *reliability coefficient* of the test for the population of individuals to be

$$\rho_t^2 = 1 - \frac{\underset{i}{E}\sigma_{t_i}^2}{\sigma_t^2} = \frac{\sigma_T^2}{\sigma_t^2}. \tag{11}$$

This definition states in precise terms what seems to be Spearman's original intention. That it is equivalent to the "correlation between two independent trials" will be seen in §10 below.

Obviously $0 \leq \rho_t^2 \leq 1$. The coefficient is zero only if all* expected scores are equal. Hence, any test which has any variance at all in expected scores has some reliability.

6. *A Comparison with the Conventional Formulation.* The term $\underset{i}{E}\sigma_{t_i}^2$ in (10) is equivalent to what has been called the "error variance" in previous formulations. Notice, however, the precise analytical structure of this term: *it is the mean of the individual error variances.* No assumption is made that individuals are equally unreliable. Whatever individual differences there may be in unreliability, the error term is the *mean* of the unreliability variances.

The term σ_T^2 corresponds to what has been called the "true variance" in previous formulations. Notice that it is not assumed that

$$\underset{i}{E}(t_{ik} - T_i) = 0, \tag{a}$$

that is, that the mean of the errors of a single trial is zero, as is done conventionally. Nor is it assumed that the errors correlate zero with the true scores on a single trial:

$$\underset{i}{E}(t_{ik} - T_i)(T_i - \mu_T) = 0, \tag{b}$$

as is done conventionally. Formula (10) actually *involves no assumptions* except that the variances exist.† Therefore, definition (11) defines the reliability coefficient *for dependent trials as well as for independent trials.*

Concern about independence of trials arises from the need to set bounds for ρ_t^2 on the basis of one or two trials. Parts II and III of this paper show how such bounds can be set, given independence between items and between persons over trials.

7. *The Definition of Experimental Independence.* The definition of *experimental* independence involves the notion of *statistical* independence. The general definition of the *statistical* independence of two variates, continuous or discrete, is well known, but it may be

* More strictly, all except possibly for an infinitesimal proportion.

† That (a) and (b) are actually true, except possibly for an infinitesimal proportion of trials, can be proved using an assumption of independent trials, using the method of Part III.

helpful to review it here. Let y be a variate with frequency function $f_1(y)$; let z be a variate with frequency function $f_2(z)$; let $f(y,z)$ be the joint frequency function of y and z; let $g(y|z)$ be the conditional distribution of y for fixed z; and let $h(z|y)$ be the conditional distribution of z for fixed y. Then y is said to be statistically independent of z if

$$g(y|z) = f_1(y) \tag{c}$$

for each z. Similarly, z is statistically independent of y if

$$h(z|y) = f_2(z) \tag{d}$$

for each y. It is well known that if (c) is true, then (d) is true, and conversely; and that a necessary and sufficient condition for (c) and (d) is that

$$f(y,z) = f_1(y) f_2(z). \tag{e}$$

As a consequence of (e), it follows that, for any powers p and q for which the indicated moments exist,

$$Ey^p z^q = (Ey^p)(Ez^q), \tag{f}$$

where the expectations are taken over the appropriate universes.

In the present paper, we find it convenient to consider a distribution of observations as an unarranged sequence of numbers, so that we need to make no explicit statements in terms of frequency functions. At most we shall require conditions analogous to (f), where the highest power is *four.* That is, instead of requiring complete independence, we require at most only such independence as is defined by the first four moments. For brevity, however, we shall at times speak of "independence" without qualification, even though it is more than we need.

By *experimental independence* we mean *statistical independence in a universe of trials.* As an example, let y_{ik} and z_{ik} be the values of y and z of the ith person on the kth trial. If y and z are experimentally independent for the ith person, then it must be true for such p and q for which the moments exist that

$$\underset{k}{E} y_{ik}^p z_{ik}^q = (\underset{k}{E} y_{ik}^p)(\underset{k}{E} z_{ik}^q). \tag{g}$$

Another way of stating (g) which is more convenient is:

$$\underset{k}{E}(y_{ik}^p - \underset{k}{E} y_{ik}^p)(z_{ik}^q - \underset{k}{E} z_{ik}^q) = 0. \tag{g$'$}$$

In particular, when $p = q = 1$, (g') states that the *covariance* over trials between y and z is zero for the ith person.

Notice that (g) expresses an independence condition *for each person separately.* For brevity, we shall state that two variates are experimentally independent without inserting a qualification, thereby implying that this is true for each individual.

It should be clear that the definition of experimental independence does *not* involve the relation between y and z considered over the population of people. For example, (g) does not involve $\underset{i}{E} y_{ik}{}^p z_{ik}{}^q$.

The extension of the notion of experimental independence to more than two variables follows easily. An explicit statement of what is needed for this paper is given in assumption (C), in the next section.

8. *The Basic Assumptions.* There are only three kinds of assumptions employed in this paper:

Assumption (A). The following moments exist:

$$\underset{i}{E}\underset{k}{E} x_{ijk}{}^p \quad (p = 1, 2, 3, 4; j = 1, 2, \cdots, n),$$

where expectations are taken first over k and then over i.

Assumption (B). The population of individuals and the universe of trials are indefinitely large.

Assumption (C). The items are experimentally independent to the extent defined by the equations

$$\underset{k}{E}(x_{h_1gk} - X_{h_1g})^p (x_{i_1gk} - X_{i_1g})^q (x_{h_2jk} - X_{h_2j})^r (x_{i_2jk} - X_{i_2j})^s$$
$$= [\underset{k}{E}(x_{h_1jk} - X_{h_1j})^p] [\underset{k}{E}(x_{i_1jk} - X_{i_1j})^q] [\underset{k}{E}(x_{h_2gk} - X_{h_2g})^r] [\underset{k}{E}(x_{i_2gk} - X_{i_2g})^s],$$

$h_1 \neq i_1$; $h_2 \neq i_2$; if $\underline{g} = j$, then $h_1 \neq h_2$ or i_2, and $i_1 \neq i_2$ or h_2; $p, q, r, s = 0, 1, 2, \underline{3; p + q + r + s \leqq} 4$; $h_1, i_1, h_2, i_2 = 1, 2, \cdots$; $g, j = 1, 2, \cdots, n$.

Let us examine how stringent these assumptions are.

From the theory of moments, it is well known that if a distribution has a finite range, then all its moments exist. Since test items ordinarily permit only finite scores, assumption (A) is in practice almost invariably fulfilled. If assumption (A) is fulfilled, it follows that the moments

$$\underset{i}{E}\underset{k}{E} t_{ik}{}^p \quad (p = 1, 2, 3, 4)$$

exist, since t_{ik} is the sum of a finite number of items.*

Thus far in this paper we have assumed only that $\underset{i}{E}\underset{k}{E}t_{ik}$ and $\underset{i}{E}\underset{k}{E}t^2_{ik}$ exist, which is enough to establish basic equation (10), from which definition (11) was formed.

Assumption (B) is not needed until Part III, where it serves as part of the sufficient conditions for lower bounds to ρ_t^2 to be observable from a single trial. That the universe of trials be indefinitely large seems part of the *definition* of the problem of test-retest reliability, rather than an empirical assumption; errors of observation seem always to be regarded as a sample from some hypothetical universe of indefinitely many trials. That the population of *individuals* be indefinitely large, on the other hand, is more of an empirical restriction. There are cases where it may be desired to speak of the reliability of a test for a given, relatively small, group of people. The biases in setting a lower bound from only a single trial in such a case are shown in the results of Part III. These biases, however, disappear for indefinitely large populations.

While explicit use is made in Part III of the assumption that the population of individuals is indefinitely large, no explicit statement is made of the corresponding assumption for the universe of trials. In fact, nowhere in any derivations of this paper is it explicitly stated that the trials are indefinitely numerous. We therefore point out here that it is *implicit* in the derivations that if there are N persons in the population, then there must be at least $N + 1$ trials in the universe if two items are to be experimentally independent. This is because assumption (C) imposes at least $N + 1$ linear restrictions on the x_{ijk}, so that if the population is indefinitely large, assumption (C) cannot hold unless the universe of trials is also indefinitely large.

9. *The Assumptions of Experimental Independence.* Assumption (C) states essentially two things:

(C_1) The observed value of an individual on an item is experimentally independent of his values on any other items.

(C_2) The observed value of an individual on an item is experimentally independent of the observed values of any other individual on that or any other item.

Assumption (C_2) is used in Part III of this paper. Experimental

* It becomes another problem to consider limiting cases as $n \rightarrow \infty$, since limits of the moments of t_{ik} need not exist. Consideration of such limiting cases is required if inferences are sought concerning a *universe of items* of which the n are a sample. Ordinarily, this can be done by taking *expectations* over the universe of items, rather than sums, but for this the universe of items should have some specified structure. A scale (2) is a simple example of a structured universe of items.

conditions can usually be established to fulfill this assumption. For example, if the set of items is an objective examination or an attitude questionnaire, and if the individuals cannot copy from each other and can have no indication of how the others are answering, their responses may ordinarily be considered experimentally independent.

Assumption (C_1) is used more immediately, in Part II, and may afford more difficulty in being realized in practice. What it calls for is that the deviations from trial to trial of the values of an individual on one item shall not depend on the deviations on the other items. In an examination or questionnaire of any substantial length, this condition can often be approached by an appropriate arrangement of the items into such a sequence as to minimize carry-over from item to item within a trial.

It should be remarked that nowhere is it assumed that items are "equivalent," or are "measuring the same thing." The items may be a battery of heterogeneous predictors, assembled to predict some one criterion; they may be a sample from a scalable universe, and thus be essentially functions of a single variable (2); or they may be anything else. No assumptions are needed beyond (A), (B), and (C) to make a practical investigation of reliability. Further assumptions like those used conventionally are unnecessary, and are indeed impractical since ordinarily it is difficult to tell when—if at all— they are fulfilled.

10. *Correlation Between Two Trials.* Before proceeding to our new results, it may be of interest to see what a classical result looks like in our present terms. We shall show that ρ_t^2 is the same as the correlation between two independent trials, when such a correlation is properly defined. In order to speak of two trials as experimentally independent, we must regard them as one pair out of a *universe of pairs of trials.*

Let t_{1ik} and t_{2ik} be the test scores of the ith person on the kth pair of trials. We have, by definition,

$$\begin{gathered}\underset{k}{E}\,t_{1ik} = \underset{k}{E}\,t_{2ik} = T_i \\ \underset{i\ k}{EE}\,t_{1ik} = \underset{i\ k}{EE}\,t_{2ik} = \underset{i}{E}\,T_i = \mu_t \\ \underset{i\ k}{EE}\,(t_{1ik} - \mu_t)^2 = \underset{i\ k}{EE}\,(t_{2ik} - \mu_t)^2 = \sigma_t^2 .\end{gathered} \tag{12}$$

The covariance between pairs over all individuals and pairs of trials is

$$\gamma_{t_1 t_2} = \underset{i\ k}{EE}\,(t_{1ik} - \mu_t)\,(t_{2ik} - \mu_t)\,, \tag{13}$$

and the correlation between pairs over all individuals and pairs of trials is clearly

$$\rho_{t_1t_2} = \frac{\gamma_{t_1t_2}}{\sigma_t^2}. \tag{14}$$

Now, we assume that the trials are experimentally independent at least to the extent that covariances over trials vanish for each person:

$$\underset{k}{E}(t_{1ik} - T_i)(t_{2ik} - T_i) = 0, \quad (i = 1, 2, \cdots). \tag{15}$$

Rewriting (13) as

$$\begin{aligned}\gamma_{t_1t_2} = \underset{i}{E}\underset{k}{E}[(t_{1ik} - T_i) \\ + (T_i - \mu_T)][(t_{2ik} - T_i) + (T_i - \mu_T)],\end{aligned} \tag{16}$$

expanding, using (15), (12), and (8), we obtain

$$\gamma_{t_1t_2} = \sigma_T^2. \tag{17}$$

From (17), (14), and (11),

$$\rho_{t_1t_2} = \rho_t^2. \tag{18}$$

It should be carefully noted that (14) does *not* define the correlation over individuals between test scores on *two trials*; it defines the correlation over individuals and over a *universe of pairs of trials*. The correlation over individuals for a pair of trials is defined by

$$\rho_{t_1t_2,k} = \frac{\underset{i}{E}(t_{1ik} - \underset{i}{E}t_{1ik})(t_{2ik} - \underset{i}{E}t_{2ik})}{\sqrt{[\underset{i}{E}(t_{1ik} - \underset{i}{E}t_{1ik})^2][\underset{i}{E}(t_{2ik} - \underset{i}{E}t_{2ik})^2]}}. \tag{19}$$

It can be shown, using assumptions (A) and (B), and restating (C) to refer to t_{1ik} and t_{2ik}, that except possibly for an infinitesimal proportion of the time, $\rho_{t_1t_2,k}$ is equal to $\rho_{t_1t_2}$, and hence to ρ_t^2. This can be seen by analogy to the results of Part III.

Therefore, if it is possible to make two independent trials of a test in practice, on a *large* population, then by virtue of Part III, the correlation between the two trials may be taken as equal to the reliability coefficient.

PART II. *Lower Bounds for the Reliability Coefficient*

11. *The Item Variances and Covariarces.* Because of the empirical difficulties of making two independent trials, it becomes im-

portant to inquire into what can be learned about reliability from only a single trial. It is to this problem that the remainder of this paper is devoted. A single trial can yield information about the reliability of the whole test if at least two experimentally independent items are involved. In such a case, lower bounds to the reliability coefficient can be computed from item variances and covariances observed on a single trial.

To establish a lower bound, we first derive it in terms of parameters defined over all individuals and trials. It is then shown in Part III that the required parameters are observable from but a single trial.

Let the total mean of the jth item be μ_j. Then

$$\mu_j = \underset{i}{E}\,\underset{k}{E}\,x_{ijk} = \underset{i}{E}\,X_{ij} \qquad (j = 1, 2, \cdots, n). \qquad (20)$$

The variance of the expected scores and the total variance, respectively, of the jth item are

$$\sigma^2_{X_j} = \underset{i}{E}\,(X_{ij} - \mu_j)^2 \qquad (j = 1, 2, \cdots, n) \qquad (21)$$

$$\sigma^2_{x_j} = \underset{i}{E}\,\underset{k}{E}\,(x_{ijk} - \mu_j)^2 \qquad (j = 1, 2, \cdots, n). \qquad (22)$$

The covariance between expected scores and the total covariance, respectively, for items g and j are

$$\gamma_{X_gX_j} = \underset{i}{E}\,(X_{ig} - \mu_g)(X_{ij} - \mu_j) \qquad (g, j = 1, 2, \cdots, n) \qquad (23)$$

$$\gamma_{x_gx_j} = \underset{i}{E}\,\underset{k}{E}\,(x_{igk} - \mu_g)(x_{ijk} - \mu_j), \quad (g, j = 1, 2, \cdots, n). \qquad (24)$$

As usual, the covariance of a variable with itself is its variance:

$$\gamma_{X_jX_j} = \sigma^2_{X_j}; \quad \gamma_{x_jx_j} = \sigma^2_{x_j}.$$

Rewriting (24) in the form

$$\gamma_{x_gx_j} = \underset{i}{E}\,\underset{k}{E}\,[(x_{igk} - X_{ig}) + (X_{ig} - \mu_g)][(x_{ijk} - X_{ij}) + (X_{ij} - \mu_j)],$$

expanding, using (2) and assumption* (C), we obtain a basic formula:

$$\gamma_{x_gx_j} = \gamma_{X_gX_j} + \delta_{gj}\,\underset{i}{E}\,\sigma^2_{x_{ij}}, \qquad (g, j = 1, 2, \cdots, n), \qquad (25)$$

where δ_{gj} is Kronecker's delta:

$$\delta_{gj} = \begin{cases} 1 & g = j \\ 0 & g \neq j \end{cases}.$$

* Actually only (C_1), and even then only with regard to covariances.

The two kinds of statements in (25) are explicitly:

$$\gamma_{x_g x_j} = \gamma_{X_g X_j}, \qquad g \neq j \tag{26}$$

$$\sigma^2_{x_j} = \sigma^2_{X_j} + \underset{i}{E}\, \sigma^2_{x_{ij}} . \tag{27}$$

In Part III, it is shown that the $\gamma_{x_g x_j}$, including the $\sigma^2_{x_j}$, are observable from a single trial. Item covariances from a single trial are defined by

$$\gamma_{gj.k} = E\,(x_{igk} - \underset{i}{E} x_{igk})\,(x_{ijk} - \underset{i}{E} x_{ijk}) .$$

In §20, it is shown that $\gamma_{gj.k} = \gamma_{x_g x_j}$, except possibly for an infinitesimal proportion of the trials. Therefore, from (26) the $\gamma_{X_g X_j}$ are also observable when $g \neq j$. But the $\sigma^2_{X_j}$ are in general not observable, because error variances $\underset{i}{E}\, \sigma^2_{x_{ij}}$ are not observable. Similarly, $\sigma_t{}^2$ is observable, but $\sigma_T{}^2$ is not.

The lower bounds developed in the following sections are entirely in terms of observable quantities.

12. *A Simple Lower Bound.* Subtracting corresponding members of (5) from (3), we have

$$t_{ik} - T_i = \sum_{j=1}^{n} (x_{ijk} - X_{ij}) .$$

Squaring both members, taking expectations over k, and using (6), (2), and assumption (C_1), yields

$$\sigma^2_{t_i} = \sum_{j=1}^{n} \sigma^2_{x_{ij}} . \tag{28}$$

This states that the *test error variance of a person is the sum of his item error variances.* Taking expectations over i in (28) yields

$$\underset{i}{E}\, \sigma^2_{t_i} = \sum_{j=1}^{n} \underset{i}{E}\, \sigma^2_{x_{ij}} . \tag{29}$$

This states that the *total error variance of the test is the sum of the item error variances.*

Summing both members of (27) over j and using (29) yields another basic formula:

$$\sum_{j=1}^{n} \sigma^2_{x_j} = \sum_{j=1}^{n} \sigma^2_{X_j} + \underset{i}{E}\, \sigma^2_{t_i} . \tag{30}$$

The left member of (30) is observable from a single trial, so we immediately have a useful and simple inequality:

$$\sum_{j=1}^{n} \sigma^2_{x_j} \geqq \underset{i}{E}\, \sigma^2_{t_i} \,. \tag{31}$$

Using this in (11), we obtain a simple lower bound to the reliability coefficient that is observable on a single trial. Let

$$\lambda_1 = 1 - \frac{\sum_{j=1}^{n} \sigma^2_{x_j}}{\sigma_t^2} . \tag{32}$$

Then

$$\lambda_1 \leqq \rho_t^2 \leqq 1 \,. \tag{33}$$

The equality on the left holds if and only if $\sigma^2_{x_j} = 0$, $(j = 1, 2, \cdots, n)$, in which case $\sigma_T^2 = \rho_t^2 = 0$.

From (3), (9), and (24), it readily follows that

$$\sigma_t^2 = \sum_{g=1}^{n} \sum_{j=1}^{n} \gamma_{x_g x_j} .$$

Then we can rewrite the lower bound as

$$\lambda_1 = \frac{\sigma_t^2 - \sum_{j=1} \sigma^2_{x_j}}{\sigma_t^2} = \frac{\sum_{g=1}^{n} \sum_{j=1}^{n} \gamma_{x_g x_j} - \sum_{j=1}^{n} \sigma^2_{x_j}}{\sum_{g=1}^{n} \sum_{j=1}^{n} \gamma_{x_g x_j}} . \tag{34}$$

The right member shows that the numerator is the sum of all the observed item covariances, omitting the variances, while the denominator is the sum of all the variances and covariances. Let

$$\Gamma_1 = \sigma_t^2 - \sum_{j=1}^{n} \sigma^2_{x_j} \,. \tag{35}$$

Then λ_1 will be positive, zero, or negative according as Γ_1 is positive, zero, or negative. If λ_1 is zero or negative, then it affords no information about ρ_t^2, for ρ_t^2 is always nonnegative.

13. *A Better Lower Bound.* A better lower bound than λ_1 will now be derived. Since the square of a correlation coefficient cannot exceed unity,

$$\gamma^2_{x_g x_j} \leqq \sigma^2_{x_g} \sigma^2_{x_j} , \qquad (g, j = 1, 2, \cdots, n) \,. \tag{36}$$

Then, from (26) and (36),

$$\gamma^2_{x_g x_j} \leqq \sigma^2_{X_g} \sigma^2_{X_j}\,, \quad g \neq j, \quad (g, j = 1, 2, \cdots, n). \tag{37}$$

Summing both members over g, but subtracting out the case $g=j$, we obtain

$$\sum_{g=1}^{n} \gamma^2_{x_g x_j} - \sigma^4_{x_j} \leqq \left(\sum_{g=1}^{n} \sigma^2_{X_g} - \sigma^2_{X_j} \right) \sigma^2_{X_j}\,, \quad (j = 1, 2, \cdots, n). \tag{38}$$

Summing over all values of j, we obtain

$$\Gamma_2 \leqq \left(\sum_{j=1}^{n} \sigma^2_{X_j} \right)^2 - \sum_{j=1}^{n} \sigma^4_{X_j}\,, \tag{39}$$

where

$$\Gamma_2 = \sum_{j=1}^{n} \sum_{g=1}^{n} \gamma^2_{x_g x_j} - \sum_{j=1}^{n} \sigma^4_{x_j}\,. \tag{40}$$

Γ_2 is the sum of the *squares* of the covariances, omitting the variances. The variance of the $\sigma^2_{X_j}$ over the n items is

$$\alpha^2 = \frac{\sum_{j=1}^{n} \sigma^4_{X_j}}{n} - \left(\frac{\sum_{j=1}^{n} \sigma^2_{X_j}}{n} \right)^2, \tag{41}$$

and

$$\sum_{j=1}^{n} \sigma^4_{X_j} = n\alpha^2 + \frac{1}{n} \left(\sum_{j=1}^{n} \sigma^2_{X_j} \right)^2.$$

Substituting the right member into (39), we obtain

$$\Gamma_2 \leqq \frac{n-1}{n} \left(\sum_{j=1}^{n} \sigma^2_{X_j} \right)^2 - n\alpha^2,$$

so that clearly

$$\Gamma_2 \leqq \frac{n-1}{n} \left(\sum_{j=1}^{n} \sigma^2_{X_j} \right)^2,$$

or

$$\sqrt{\frac{n}{n-1} \Gamma_2} \leqq \sum_{j=1}^{n} \sigma^2_{X_j}\,. \tag{42}$$

This last inequality is weaker the more variation there is among the $\sigma^2_{X_j}$; in particular the equality cannot hold if $\alpha^2 > 0$.

From (30) and (42),

$$\sum_{j=1}^{n} \sigma^2_{x_j} - \sqrt{\frac{n}{n-1}\,\Gamma_2} \geqq \underset{i}{E}\, \sigma^2_{t_i}\,. \tag{43}$$

Let

$$\lambda_2 = 1 - \frac{\sum_{j=1}^{n} \sigma_{x_j}{}^2 - \sqrt{\frac{n}{n-1}\,\Gamma_2}}{\sigma_t{}^2}. \tag{44}$$

The right member is observable from a single trial. From (11) and (43), λ_2 is a lower bound to the reliability coefficient:

$$\lambda_2 \leqq \rho_t{}^2 \leqq 1\,.$$

The equality on the left holds if and only if the variances and covariances of expected scores are all equal in absolute value.

That λ_2 is a better bound than λ_1 can be seen from the fact that

$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{n}{n-1}\,\Gamma_2}}{\sigma_t{}^2}. \tag{45}$$

Another way of writing λ_2 is

$$\lambda_2 = \frac{\Gamma_1 + \sqrt{\frac{n}{n-1}\,\Gamma_2}}{\sigma_t{}^2}. \tag{46}$$

The numerator involves the sum of the covariances and the sum of the squares of the covariances, omitting the variances and their squares.

14. *An Intermediate Lower Bound.* By weakening λ_2, we can obtain a lower bound that will be simpler to compute. Let β^2 be the variance of the $n(n—1)$ covariances between items; thus,

$$\beta^2 = \frac{\Gamma_2}{n(n-1)} - \left[\frac{\Gamma_1}{n(n-1)}\right]^2. \tag{47}$$

Therefore

$$\Gamma_2 = n(n-1)\beta^2 + \frac{\Gamma_1^2}{n(n-1)},$$

or

$$\Gamma_2 \geqq \frac{\Gamma_1^2}{n(n-1)}. \tag{48}$$

The equality holds if and only if $\beta^2 = 0$, or all inter-item covariances are equal.

From (48),

$$\sqrt{\frac{n}{n-1}\Gamma_2} \geqq \frac{1}{n-1}\,|\,\Gamma_1\,|. \tag{49}$$

Using this result in (45), and remembering (34) and (35), we obtain

$$\lambda_2 \geqq \lambda_1 + \frac{1}{n-1}\,|\,\lambda_1\,|. \tag{50}$$

Therefore the right member is also a lower bound to ρ_t^2 and is intermediate between λ_1 and λ_2. But if λ_1 is negative, then the right member of (50) becomes

$$\frac{n-2}{n-1}\lambda_1,$$

which is also negative; and like λ_1, it yields no information about ρ_t^2. There is a gain in information only if λ_1 is positive. Hence, if we let the new lower bound be

$$\lambda_3 = \lambda_1 + \frac{\lambda_1}{n-1} = \frac{n}{n-1}\lambda_1, \tag{51}$$

we have lost no information by discarding the absolute value sign. Both λ_3 and λ_1 are useful if and only if λ_1 is positive.

We thus have

$$\lambda_3 = \frac{n}{n-1}\left(1 - \frac{\sum\limits_{j=1}^{n} \sigma_{x_j}^2}{\sigma_t^2}\right), \tag{52}$$

and

$$\lambda_3 \leqq \rho_t^2 \leqq 1.$$

The equality on the left holds if and only if the variances and covariances of the expected scores on the items are *all equal*, because λ_3 equals λ_2 only if $\beta^2 = 0$. (For ρ_t^2 to equal λ_2, it is necessary and sufficient that the $\gamma_{x_g x_j}$ be equal only in absolute value.)

λ_3 is easier to compute than λ_2, since only the total variance and the item variances are required.* If the covariances are all positive

* λ_3 resembles a formula developed separately by Kuder and Richardson (5) and Hoyt (4). In fact, L_3 of §2 above is algebraically identical to this formula (which is formula (20) in Kuder and Richardson's paper). This seems to be only a coincidence. The derivation of λ_3 has little in common with the derivation of

and homogeneous, then λ_3 will not be much less than λ_2 and may be an adequate lower bound. If the covariances are heterogeneous, and in particular, if some are negative, then λ_2 will be definitely superior to λ_3. λ_2 can be positive and useful when λ_3 is negative and useless.

15. *"Split-Half" Lower Bounds.* A traditional approach to estimating ρ_t^2 has been to divide a test into two parts, assume the parts to be equivalent and experimentally independent, correlate the two parts, and "correct" the correlation for test length (1, p. 419). It is fortunate that we can now dispense with assumptions of equivalence, yet retain essentially the same computations, and arrive at a rigorous answer.

If a test is divided into two parts, each part may be thought of as an item, so that the test may be thought of as composed of two items. Let σ_1^2 and σ_2^2 be the respective variances of the two parts, taken over all persons and trials. If the two parts are experimentally independent, then according to (52), with n=2, we have a "split half" lower bound to ρ_t^2. If we let

$$\lambda_4 = 2\left(1 - \frac{\sigma_1^2 + \sigma_2^2}{\sigma_t^2} \right) , \tag{53}$$

then

$$\lambda_4 \leqq \rho_t^2 \leqq 1 .$$

Let us compare λ_4 with the traditional "corrected split-half coefficient." As is readily seen,

Kuder and Richardson's formula, and has quite a different interpretation. The parameters of λ_3 are defined over trials and persons; λ_3 is a lower bound to ρ_t^2; and the proof that λ_3 can be computed from but a single trial by L_3 requires no assumption beyond (C). Kuder and Richardson's approach makes no treatment of a universe of trials; it attempts to estimate rather than to bound the reliability coefficient; it introduces assumptions about the relationships between items; and observability from but a single trial is not explicitly discussed. Hoyt's analysis does seem to consider a universe of trials, albeit observability from a single trial is not explicitly examined; however, he introduces very stringent hypotheses about the relationships between items (which in a sense resemble Kuder and Richardson's) in terms of a linear hypothesis of analysis of variance; he attempts to estimate rather than to bound the reliability coefficient. The stringency of this linear hypothesis can be seen from the fact that, in our notation, it requires the equality in (36) to hold for all pairs of items and that $\alpha = \beta = 0$. But these conditions are rarely, if ever, found in practice. The hypothesis that $\beta = 0$ can be tested from a single trial. If the items are part of an approximate scale (2), it can be seen that $\beta \neq 0$; and if the items are a heterogeneous set of predictors for a particular criterion, then ordinarily $\beta \neq 0$. The hypotheses that the equality in (36) holds and that $\alpha = 0$ cannot be tested from but a single trial, which makes it desirable to avoid them, especially in view of their strictness.

λ_3 is simple and should be very useful, even though it is weaker than λ_2. It is indeed fortunate that L_3 needs no assumption beyond assumption (C) to be used as a lower bound to the reliability coefficient in practice.

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 + 2\,\sigma_1\,\sigma_2\,\rho_{12}\,, \tag{54}$$

where ρ_{12} is the correlation between the two parts over all persons and trials. If $\sigma_1^2 = \sigma_2^2$, then from (54) and (53), we obtain

$$\lambda'_4 = \frac{2\,\rho_{12}}{1 + \rho_{12}}\,, \tag{53'}$$

which resembles the "corrected split-half" formula.

But λ'_4 is still a *lower bound* to ρ_t^2. It does *not* assume equivalent halves—it assumes only experimentally independent halves with equal variances. If the two halves are not equivalent, (but have equal variances) then λ'_4 *definitely underestimates* ρ_t^2.

This has a very important bearing on all previous empirical research which has used "corrected split-half coefficients." It seems plausible that in a vast number of cases the hypothesis of equivalence is dubious, and that therefore *the reliability of a great many tests has been seriously underestimated.*

If reliability is underestimated, *attenuation is overestimated.* Therefore, "corrections for attenuation" used in past research (1, p. 367) must be regarded with caution, for in many cases they are fallaciously high. This helps account for "corrections" that yield correlations greater than unity in practice.

In practice, there is no need to use even the hypothesis $\sigma_1^2 = \sigma_2^2$. Formula (53) is preferable to (53′) because it needs no such assumption.

If a test contains many items, there are many ways of splitting it into two parts, and each way may yield a different λ_4. The largest of these λ_4 is, of course, the best bound for ρ_t^2 computed in this manner. In practice, a sufficiently high λ_4 can often be found by a careful splitting of the test, so that there would be no need to seek the best λ_4.

If a test can be split into two parts that correlate relatively highly with each other, this will ordinarily yield a λ_4 that is better than λ_3 or even λ_2.

16. *A Lower Bound Based on a Best Row of Covariances.* A lower bound which may in some cases be better than λ_2 can be established as follows. From (38), we see that

$$\sum_{g=1}^{n} \sigma^2_{X_g} \geqq \sigma^2_{X_j} + \frac{1}{\sigma^2_{X_j}}\,\Gamma_{2j}\,, \qquad (j = 1, 2, \cdots, n)\,, \tag{55}$$

where

$$\Gamma_{2j} = \sum_{g=1}^{n} \gamma^2_{x_g x_j} - \sigma^4_{x_j}\,, \qquad (j = 1, 2, \cdots, n)\,. \tag{56}$$

By differentiation, the minimum of the right member of (55) as a function of $\sigma^2_{x_j}$ is found to be attained when $\sigma^2_{x_j}$ is equal to $\sqrt{\Gamma_{2j}}$. Using this minimizing value, we obtain

$$\sum_{g=1}^{n} \sigma^2_{x_g} \geqq 2\sqrt{\Gamma_{2j}}, \qquad (j = 1, 2, \cdots, n). \tag{57}$$

Let

$$\lambda_{5j} = 1 - \frac{\sum_{g=1}^{n} \sigma^2_{x_g} - 2\sqrt{\Gamma_{2j}}}{\sigma_t^2}, \qquad (j = 1, 2, \cdots, n). \tag{58}$$

From (30), (11), and (58) we see that we have n lower bounds to ρ_t^2:

$$\lambda_{5j} \leqq \rho_t^2 \leqq 1, \qquad (j = 1, 2, \cdots, n). \tag{59}$$

Let λ_5 be the largest of the λ_{5j}. Then λ_5 is the best lower bound based on the covariances with a single item in this fashion. Explicitly,

$$\lambda_5 = 1 - \frac{\sum_{g=1}^{n} \sigma^2_{x_g} - 2\sqrt{\bar{\Gamma}_2}}{\sigma_t^2} \tag{60}$$

where $\bar{\Gamma}_2$ is the largest of the Γ_{2j}.

That λ_5 is in general better than λ_1 follows from the fact that

$$\lambda_5 = \lambda_1 + \frac{2\sqrt{\bar{\Gamma}_2}}{\sigma_t^2}.$$

λ_5 will be greater than λ_2 in those cases where

$$2\sqrt{\bar{\Gamma}_2} > \sqrt{\frac{n}{n-1}\Gamma_2},$$

which requires that $n > 2$ and that one item have large absolute covariances with the other items compared with the covariances among those items. Otherwise, $\lambda_5 \leqq \lambda_2$.

17. *A Lower Bound Based on Linear Multiple Correlation.* All the preceding lower bounds will often fail to approach unity in practice, even though $\rho_t^2 = 1$. In this section, we shall see that a high lower bound can be computed for the case where each item has a high linear multiple correlation on the remaining items.

Let β_{jg} $(j, g = 1, 2, \cdots, n)$ be any set of n^2 constants, and let

$$y_{ijk} = \sum_{g=1}^{n} \beta_{jg}(x_{igk} - \mu_g). \tag{61}$$

Thus, y_{ijk} is an arbitrary linear combination of the x_{igk} deviates. Let

$$\varepsilon_j{}^2 = \underset{i\ k}{EE}[(x_{ijk}-\mu_j)-y_{ijk}]^2, \quad (j=1,2,\cdots,n) \tag{62}$$

and let

$$\eta_j{}^2 = \underset{i\ k}{EE}[(X_{ij}-\mu_j)-y_{ijk}]^2, \quad (j=1,2,\cdots,n). \tag{63}$$

Upon expansion,

$$\varepsilon_j{}^2 = \sigma^2_{x_j} - 2\,\gamma_{x_jy_j} + \sigma^2_{y_j},$$

and, using (25),

$$\eta_j{}^2 = \sigma^2_{X_j} - 2(\gamma_{x_jy_j} - \beta_{jj}\underset{i}{E}\,\sigma^2_{x_{ij}}) + \sigma^2_{y_j}.$$

Therefore,

$$\varepsilon_j{}^2 - \eta_j{}^2 = \sigma^2_{x_j} - \sigma^2_{X_j} - 2\,\beta_{jj}\underset{i}{E}\,\sigma^2_{x_{ij}},$$

or, using (27),

$$\varepsilon_j{}^2 - \eta_j{}^2 = (1-2\,\beta_{jj})\underset{i}{E}\,\sigma^2_{x_{ij}}.$$

For the case $\beta_{jj} = 0$, we obtain the inequalities

$$\varepsilon_j{}^2 \geq \underset{i}{E}\,\sigma^2_{x_{ij}}, \qquad (j=1,2,\cdots,n). \tag{64}$$

The left member is a minimum when the β_{jg} are the regression coefficients of the x_{igk} in the regression of x_{ijk} on the remaining n—1 items, over the universe of trials and the population of individuals; and the minimum $\varepsilon_j{}^2$ is the variance of the errors of estimate from this regression. Hence the

Theorem: The unreliability error variance of an item is not greater than the linear regression error variance of that item from the n—1 *remaining items.**

As a corollary, *if an item has a perfect multiple correlation on the remaining items, it is perfectly reliable.*

Let $\bar{\varepsilon}_j{}^2$ be the multiple regression error variance of the jth item, and let

$$\lambda_6 = 1 - \frac{\sum\limits_{j=1}^{n}\bar{\varepsilon}_j{}^2}{\sigma_t{}^2}. \tag{65}$$

* The proof of this theorem is adaptable to afford another proof for a similar theorem for factor analysis, established elsewhere (3), to the effect that the square of the multiple correlation coefficient is a lower bound to the communality.

From (64), (65), (29), and (11), we see that λ_6 is a lower bound to the reliability coefficient:

$$\lambda_6 \leqq \rho_t{}^2 \leqq 1 .$$

Crudely speaking, λ_6 will be larger than λ_2 when the items have relatively low zero-order correlations and high multiple correlations; and λ_6 will be smaller than λ_2 when the zero-order correlations among the items are relatively high compared with the multiple correlations.

If the set of items is part of a scale (2), then as n increases the multiple correlation of any item on the remaining items will ordinarily approach unity. Therefore, given the experimental independence of the items, a sample scale with an appreciable number of items is ordinarily highly reliable.

PART III. *Observability from a Single Trial*

18. *Observability of Means from a Single Trial.* A crucial, and hitherto neglected, problem of test-retest reliability is: what parameters are observable in a single trial? In particular, to use the lower bounds established in this paper, we must show that they are observable. Since only means and covariances are required for these bounds, we shall restrict ourselves to showing that first and second moments are observable.

The proof involves the notion of *convergence in the mean*; specifically, we use the fact that if the variance of a variate is zero, the variate vanishes except possibly for an infinitesimal proportion of the observations. To bring out the details of the proof and thus obtain general results for a finite population, we shall begin with a finite population of N individuals. Then the operation $\underset{i}{E}$ is the same as the operation $\frac{1}{N}\sum\limits_{i=1}^{N}$, where the expectation is over individuals.

The mean of the jth item on a single trial is $\underset{i}{E}x_{ijk}$. The variance of this mean over all trials is

$$\underset{k}{E}(\underset{i}{E}x_{ijk}-\mu_j)^2=\underset{k}{E}[\underset{i}{E}(x_{ijk}-X_{ij})]^2$$

$$=\frac{1}{N^2}\sum_{h=1}^{N}\sum_{i=1}^{N}\underset{k}{E}(x_{hjk}-X_{hj})(x_{ijk}-X_{ij}) .$$

From assumption (C), the last member becomes

$$\frac{1}{N^2}\sum_{h=1}^{N}\sum_{i=1}^{N}\delta_{hi}\,\sigma^2_{x_{ij}}=\frac{1}{N}\underset{i}{E}\,\sigma^2_{x_{ij}} .$$

Therefore

$$\underset{k}{E}(\underset{i}{E}x_{ijk}-\mu_j)^2=\frac{1}{N}\underset{i}{E}\,\sigma^2_{x_{ij}}\,; \tag{66}$$

and, since the error variances are all bounded according to assumption (A),

$$\lim_{N\to\infty}\underset{k}{E}(\underset{i}{E}x_{ijk}-\mu_j)^2=0. \tag{67}$$

Therefore, for an infinite population, in all, except for possibly an infinitesimal proportion of trials, we have

$$\underset{i}{E}x_{ijk}=\mu_j\,,\qquad (j=1,2,\cdots,n)\,, \tag{68}$$

which was to be shown.

Equation (66) shows the important fact that if N is finite and if $\underset{i}{E}\,\sigma^2_{x_{ij}}>0$, then (68) does not hold for a substantial proportion of trials. This implies that in general, *for finite* N , *the mean of the errors of observation will not be zero on many trials.*

19. *On the Bias of Trial Variances and Covariances.* The covariances between items g and j on a single trial is

$$\gamma_{gj,k}=\underset{i}{E}(x_{igk}-\underset{i}{E}x_{igk})(x_{ijk}-\underset{i}{E}x_{ijk})\,. \tag{69}$$

In particular, of course, $\gamma_{jj,k}$ is the *variance* of the jth item in the kth trial. Using a familiar device of least squares, we write

$$\begin{aligned}\gamma_{gj,k}=\underset{i}{E}[(x_{igk}-X_{ig})+(X_{ig}-\mu_g)-\underset{i}{E}(x_{igk}-X_{ig})][(x_{ijk}-X_{ij})\\+(X_{ij}-\mu_j)-\underset{i}{E}(x_{ijk}-X_{ij})]\,,\end{aligned}$$

which upon expansion yields

$$\begin{aligned}\gamma_{gj,k}=\underset{i}{E}(x_{igk}-X_{ig})(x_{ijk}-X_{ij})+\underset{i}{E}(X_{ig}-\mu_g)(x_{ijk}-X_{ij})\\+\underset{i}{E}(X_{ij}-\mu_j)(x_{igk}-X_{ig})+\gamma_{X_gX_j}-\underset{h}{E}\underset{i}{E}(x_{hgk}-X_{hg})(x_{ijk}-X_{ij})\,.\end{aligned} \tag{70}$$

Taking expectations over k and using assumption (C), we obtain

$$\underset{k}{E}\,\gamma_{gj,k}=\gamma_{x_gx_j}-\frac{1}{N}\delta_{gj}\underset{i}{E}\,\sigma^2_{x_{ij}}\,. \tag{71}$$

The right member of (71) shows that *covariances between items* on a single trial are *unbiased* estimates of the $\gamma_{x_gx_j}$, whereas trial variances are biased as long as N is finite and there is unreliability. An important fact is that *for finite* N, *the bias in variances cannot be estimated from a single trial,* for the $\sigma^2_{x_{ij}}$ cannot be estimated from a single trial. For very large N, of course, the bias must be negligible.

20. *Observability of Variances and Covariances.* To show that universe variances and covariances are observable from a single trial involves much more detail than for the case of the mean in §18, but the proof is similar. We shall show that when the population is infinite, $\gamma_{gj,k}$ equals $\gamma_{x_g x_j}$ except possibly in an infinitesimal proportion of trials.

From (70) and (25),

$$\gamma_{gj,k} - \gamma_{x_g x_j} = \underset{i}{E}(x_{igk} - X_{ig})(x_{ijk} - X_{ij}) + \underset{i}{E}(X_{ig} - \mu_g)(x_{ijk} - X_{ij}) \tag{72}$$

$$+ \underset{i}{E}(X_{ij} - \mu_j)(x_{igk} - X_{ig}) - \delta_{gj}\underset{i}{E}\,\sigma^2_{x_{ij}} - \underset{h}{E}\underset{i}{E}(x_{hgk} - X_{hg})(x_{ijk} - X_{ij}).$$

Squaring both members, taking expectations over k, and using assumption (C), we obtain, after some tedious algebra,

$$\underset{k}{E}(\gamma_{gj,k} - \gamma_{x_g x_j})^2 = \frac{1}{N}\Bigg[\frac{N-2}{N}\underset{i}{E}\,\sigma^2_{x_{ig}}\sigma^2_{x_{ij}} + \frac{1}{N}(\underset{i}{E}\,\sigma^2_{x_{ig}})(\underset{i}{E}\,\sigma^2_{x_{ij}})$$

$$+ \underset{i}{E}(X_{ig} - \mu_g)^2\sigma^2_{x_{ij}} + \underset{i}{E}(X_{ij} - \mu_j)^2\sigma^2_{x_{ig}}\Bigg] + \frac{1}{N}\delta_{gj}\Bigg[\left(\frac{N-1}{N}\right)^2\underset{i}{E}\,\alpha_{4,x_{ij}}$$

$$+ \frac{4(N-1)}{N}\underset{i}{E}(X_{ij} - \mu_j)\alpha_{3,x_{ij}} + 2\underset{i}{E}(X_{ij} - \mu_j)^2\sigma^2_{x_{ij}} + \frac{2}{N}(\underset{i}{E}\,\sigma^2_{x_{ij}})^2 \tag{73}$$

$$- \left(2 - \frac{4}{N} + \frac{3}{N^2}\right)\underset{i}{E}\,\sigma^4_{x_{ij}}\Bigg],$$

where

$$\alpha_{p,x_{ij}} = \underset{k}{E}(x_{ijk} - X_{ij})^p, \qquad (p = 3, 4).$$

Therefore, since all the moments on the right of (73) are bounded according to assumption (A),

$$\lim_{N\to\infty}\underset{k}{E}(\gamma_{gj,k} - \gamma_{x_g x_j})^2 = 0. \tag{74}$$

Hence, for an infinite population of individuals,

$$\gamma_{gj,k} = \gamma_{x_g x_j}$$

except possibly for an infinitesimal proportion of times.

The structure of the right member of (73) indicates the caution to be observed with respect to relatively small N.

21. *Observability of the Lower Bounds.* We have thus far shown how, for an infinite population, means and covariances are observable. But each lower bound involves a *ratio of a combination* of covariances to σ_t^2. To discuss what happens to ratios for finite populations is rather complicated. The case of an infinite population is readily handled, however. As is to be expected, any of the lower bounds can be computed from a single trial provided the population is infinite. (Similarly, the reliability coefficient can be computed from two independent trials if the population is infinite.)

The proof follows immediately from combinatorial probability considerations. If each variate in a set has probability zero of *not* being constant, then the probability that at least one is not constant cannot be greater than zero. Hence the probability is unity that *all* are constant, and that any function of them is constant.

Therefore, for an infinite population, any function of the $\gamma_{gj,k}$, for fixed k, will have probability of unity of being equal to the same function of the $\gamma_{x_g x_j}$. Thus, any of the lower bounds can be computed from a single trial by substituting the observed variances and covariances of the trial for the total variances and covariances.

In conclusion, it should be emphasized that this paper has *not* concerned itself with the problem of dealing with only a sample from the population of individuals. Our results refer to a trial of all of a *large population.* Sampling problems involve far more detail, and the intricacies of the sampling distributions of ratios loom large in the picture.

REFERENCES

1. Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1936.
2. Guttman, Louis. A basis for scaling qualitative data. *American Sociological Review*, 1944, 9, 139-150.
3. Guttman, Louis. Multiple rectilinear prediction and the resolution into components. *Psychometrika*, 1940, 5, 75-99.
4. Hoyt, Cyril. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
5. Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.
6. Spearman, Charles. The proof and measurement of association between two things. *American Journal of Psychology*, 1904, 15, 72-101.
7. Spearman, Charles. Correlation calculated from faulty data. *British Journal of Psychology*, 1910, 3, 271-295.