PSYCHOMETRIKA—VOL. 8, NO. 4
DECEMBER, 1943

# THE CONCEPT OF TEST AND ITEM RELIABILITY IN RELATION TO FACTOR PATTERN

ROBERT J. WHERRY AND RICHARD H. GAYLORD
UNIVERSITY OF NORTH CAROLINA

It is shown that approaches other than the internal consistency method of estimating test reliability are either less satisfactory or lead to the same general results. The commonly attendant assumption of a single factor throughout the test items is challenged, however. The consideration of a test made up of $K$ sub-tests each composed of a different orthogonal factor disclosed that the assumption of a single factor produced an erroneous estimate of reliability with a ratio of $(n-K)/(n-1)$ to the correct estimate. Special difficulties arising from this error in application of current techniques to short tests or to test batteries are discussed. Application of this same multi-factor concept to item-analysis discloses similar difficulties in that field. The item-test coefficient approaches $\sqrt{1/K}$ as an upper limit rather than 1.00 and approaches $\sqrt{1/n}$ as a lower limit rather than .00. This latter finding accounts for an over-estimation error in the Kuder-Richardson formula (8). A new method of isolating sub-tests based upon the item-test coefficient is proposed and tentatively outlined. Either this new method or a complete factor analysis is regarded as the only proper approach to the problem of test reliability, and the item-*sub-test* coefficient is similarly recommended as the proper approach for item analysis.

The term reliability has been used loosely to apply to the resultant of the application of many different statistical operations. As a general rule these operations attempt to determine the verifiability of the original data and thus to establish the degree to which non-chance factors entered into the original measurements.

One group of techniques approaches the problem through the direct comparison of the observed distribution of measurements with that which would have arisen by chance in accordance with certain theories of probability. These techniques parallel the use of the critical ratio in establishing the reliability of means, sigmas, differences, and other statistical parameters by disproving the null hypothesis. One such approach is that of Jackson (7), who measures the sensitivity of the test, gamma ($\gamma$), by taking the ratio of the standard deviation of the capacity to the standard deviation of chance. Another writer, Hoyt (6), using the analysis of variance approach, suggests using the ratio of the true variance of the student responses to

the obtained variance among students. Edgerton and K. F. Thomson (3) suggest using the Lexis ratio to show that differences among students are greater than those among items. Hoyt shows that his results are comparable to those of Jackson, and both Hoyt and Edgerton and Thomson show their results to be comparable to those achieved by the use of certain of the Kuder-Richardson series [actually to formulas (14) and (20), which involve only the sigmas of the items and the sigma of the total test] which we will discuss later at some length. Any restrictions demonstrated to hold for these particular portions of the Kuder-Richardson series will thus apply to these probability methods as well. Allied techniques are (1) the Horst (4) maximized criterion which holds that the most reliable weighted composite is that with the largest relative variance, and (2) the item selection techniques based on the variance (difficulty) of the item. The remaining techniques attack the measurement of reliability through the verifiability of the original ranks of the members of the population. These techniques form two natural groups as the emphasis on stability is (1) regardless of time, or (2) regardless of the particular test or items used.

The first of these viewpoints—verifiability regardless of time—is exemplified by the test-retest method of measuring reliability. While appearing best to meet the operations indicated in the usual definition of reliability, this method has been widely criticised from many different viewpoints. Typical criticisms refer to the effects of differential practice, memory, inability to duplicate testing conditions, inability to sustain motivation, etc. Perhaps even more serious is the objection raised that this coefficient is affected not only by the unreliability of the test but also by the unreliability (lack of stability) of the function being tested. Paulsen (12) suggested measuring this trait fluctuation by correcting the test-retest coefficient for attenuation, using the split-half method for obtaining the reliabilities of the initial and final testings. Woodrow (18) suggested measuring this "quotidian variability" by the ratio of the actual sigma of the means of various samples to the sigma of the means as predicted from the average standard deviation of the multiple samples. Thouless (17) proposed the measurement of this "functional fluctuation" by what he called the double test-retest index, where he substituted alternate forms for the split-half approach of Paulsen. Thus in application we see that the "regardless of time criterion" in addition to many other ills requires an appeal to the preceding view—as in Woodrow—or to the remaining criterion of stability "regardless of test or items used" criterion—as in the split-half or comparable form methods of Paulsen or Thouless. Indirectly associated techniques are: (1) the original

Brown (1) concept as to the proper values for substitution in the Brown-Spearman formula; and (2) the concept of testing item reliability by means of individual changes in response on future test adminstrations.

The remaining concept—the verifiability of the measurements regardless of the particular test used—is exemplified by the "comparable test" method of measuring reliability. Here the procedure is complicated by two major difficulties: (1) the time element, and (2) the determination of *comparability*. With respect to time the argument centers in the merit of separate alternate tests (assuming "comparability" for the time being), with the necessary time separation involved, as compared to some split-half (odds vs. evens, first half-last half, etc.) technique (again assuming "comparability") which would eliminate the time element. All of the objections raised against the test-retest method with the possible exception of the memory element would also apply against any time-separated alternate form approach. Most recent writers have assumed the desirability of eliminating this time element by some method based on internal analysis.

The second difficulty involved in this last viewpoint—the nature and determination of "comparability"—has usually been dealt with inadequately. It is always assumed but rarely demonstrated. The only really adequate definition of such comparability or equivalence known to the writers is that given by Kuder and Richardson (9) as their equation (1) which we hereby adopt. They say,

"The correlation between two forms of a test is given by

$$r_{(a+b+c+\cdots+n)(A+B+C+\cdots+N)} = \frac{r_{aA}\,\sigma_a\,\sigma_A + r_{aB}\,\sigma_a\,\sigma_B + \cdots + r_{nM}\,\sigma_n\,\sigma_M + r_{nN}\,\sigma_n\,\sigma_N}{[\sigma_a^2 + \sigma_b^2 + \cdots + \sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + r_{ac}\,\sigma_a\,\sigma_c + \cdots + r_{mn}\,\sigma_m\,\sigma_n)]^{\frac{1}{2}}} \quad (1)$$
$$[\sigma_A^2 + \sigma_B^2 + \cdots + \sigma_N^2 + 2(r_{AB}\,\sigma_A\,\sigma_B + r_{AC}\,\sigma_A\,\sigma_C + \cdots + r_{MN}\,\sigma_M\,\sigma_N)]^{\frac{1}{2}}$$

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

As Kuder and Richardson point out, the above definition of comparability makes the two terms in the bottom of equation (1) identical, which reduces the formula for the true reliability of a test or test

battery with unit weights to the form

$$r_{tt} = \frac{r_{aa}\,\sigma_a^2 + r_{bb}\,\sigma_b^2 + \cdots + r_{nn}\,\sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + \cdots + r_{mn}\,\sigma_m\,\sigma_n)}{\sigma_a^2 + \sigma_b^2 + \cdots + \sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + \cdots + r_{mn}\,\sigma_m\,\sigma_n)}. \quad (2)$$

If such a test as the capital letter series above were available it would truly be an alternate or comparable form. Actually such a test is seldom if ever available. Attempts at constructing alternate forms seldom hold rigorously to the above definitions. Instead it has been the custom to make either explicitly or implicitly the assumption of a single factor running through all such possible items and to construct alternate forms paying attention only to the difficulty of the items (equating means and sigmas, if even these are taken into account).

The internal consistency hypothesis is the basis of the two most common methods of measuring reliability: (1) the split-half Spearman-Brown (15) approach, and (2) the Kuder-Richardson series. Kuder and Richardson frankly assume a single factor among the test items, while the Spearman-Brown assumption of equal intercorrelations amounts to the same thing as is demonstrated when Kuder and Richardson derive the Spearman-Brown formula at one stage in their own series. Allied techniques are (1) the Edgerton-Kolbe (2) conception of the most reliable weighted criterion based on minimal differences among the scores of each individual for each of the criteria; (2) the Hotelling (5) conception of the most predictable criterion with weights proportional to the loadings on the first unrotated principle component; (3) and the many methods of item analysis, too numerous even to list, based on the item-test correlation coefficient or variants of that measure. When the assumption of a single factor is satisfied such procedures are justified, but frequently that assumption is not justified as is indicated by the many studies which show the usual test or test battery to contain many factors. What is the result of using these formulas when, as is probably almost universally the case, the basic assumption is not justified? Have we, perhaps been branding as unreliable tests which were satisfactory in that respect? Have we been discarding as unreliable items which were perfectly good? The remainder of this paper considers these questions theoretically. Instead of the usual assumption of a single factor with equal intercorrelations and equal sigmas throughout the whole test, we shall instead assume a test of $K$ factors—$a$, $b$, $c$, $\cdots$, $k$—each factor being represented by a number of items $n_x$ which may vary from factor to factor. For the sake of simplicity and to maximize the difference between this case and the usual assumptions these factors shall be taken as orthogonal or uncorrelated, i.e., the intercorrelations

between items in different factor groups are taken to be zero.* While the sigmas and inter-item correlations within a factor group may differ from group to group ($r_{aa} \neq r_{ff}$) and ($\sigma_a \neq \sigma_f$), we shall make the usual assumptions of equal sigmas and equal intercorrelations within each factor group ($r_{a_1a_5} = r_{a_2a_7}$ and $\sigma_{a_4} = \sigma_{a_6}$, etc.) and shall further assume the reliability of each item to be equal to the inter-item correlation holding for items in that group. Substituting these assumptions in equation (2) we see that the reliability of this multiple-factor test would be

$$r_{tt} = \frac{\sum^k n_x^2 \sigma_x^2 r_{xx}}{\sum^k n_x \sigma_x^2 + \sum^k n_x(n_x - 1) \sigma_x^2 r_{xx}}. \tag{3}$$

We want to compare with this value, the value as estimated by the single factor theory. The Kuder-Richardson formula (14) is taken as the best measure of this type since it is this formula which the various direct probability approaches equalled and which Kuder and Richardson have shown to be equal to the Brown-Spearman method which is basic to split-half approaches to the measurement of reliability. We shall also see later on that this is the most general of their formulas which is applicable, since formula (8) of their series is erroneous. The Kuder-Richardson formula (14) reads

$$r_{tt_{KR_{14}}} = \frac{\sigma_t^2 - \sum^n \sigma_x^2}{(\sum^n \sigma_x)^2 - \sum^n \sigma_x^2} \cdot \frac{(\sum^n \sigma_x)^2}{\sigma_t^2}. \tag{4}$$

Substituting our present assumptions for our $K$-factored test in this equation gives

$$r_{tt_{KR_{14}}} = \frac{\sum^K n_x(n_x - 1) \sigma_x^2 r_{xx}}{(\sum^K n_x \sigma_x)^2 - \sum^K n_x \sigma_x^2} \cdot \frac{(\sum^K n_x \sigma_x)^2}{\sum^K n_x \sigma_x^2 + \sum^K n_x(n_x - 1) \sigma_x^2 r_{xx}} \tag{5}$$

and to bring out the difference between equation (3) and (5) more clearly we can rewrite (5) as

* One of the editors objected to the stringency of this case, on the basis that it would seldom if ever occur in practice. We deliberately chose to magnify the discrepancy between our results and those based upon the assumption of a single factor, due to our feeling that the actual usual case would lie somewhere in between the two extremes. The special case formulas derived in this paper were not meant to be used computationally but only to disclose the effects of various possible trends which might exist in practical situations. We do believe that the possible sub-tests and their inter-relations form the only sound approach to either the problem of test reliability or the problem of item-analysis.

$$r_{tt_{KR14}} = \frac{(\sum^K n_x \sigma_x)^2}{(\sum^K n_x \sigma_x)^2 - \sum^K n_x \sigma_x^2} \left[ \frac{\sum^K n_x^2 \sigma_x^2 r_{xx}}{\sum^K n_x \sigma_x^2 + \sum^K n_x (n_x - 1) \sigma_x^2 r_{xx}} \right. \tag{5a}$$
$$\left. - \frac{\sum^K n_x \sigma_x^2 r_{xx}}{\sum^K n_x \sigma_x^2 + \sum^K n_x (n_x - 1) \sigma_x^2 r_{xx}} \right] .$$

Now the first term in this bracket is the true value as given by equation (3), which we see is reduced by an error factor (the second term in the bracket) and then increased again by a multiplier greater than one (the term preceding the bracket). We can see that the extent of this error will depend upon the actual values of the $n$'s, $\sigma$'s, and $r$'s. To evaluate the extent of this error under certain special conditions we can simplify the formulas by assuming various ones of these determiners, $n_x$, $\sigma$, and $r$, to be equal from group to group. The following table gives the reduced form for $r_{tt}$ and $r_{ttKR14}$ for each such possible set of assumptions.

Several conclusions of interest and importance can be drawn on the basis of the equations in set (6). We note that:

(a) The Kuder-Richardson formula equals the true formula only when $K$, the number of factors, equals one (Case g, since for the remaining single set the $n$'s, $\sigma$'s and $r$'s would all be equal, with $K$ equal to one). With $K$ equal to one, both formulas take the usual Brown-Spearman form, indicating that that formula is also correct for the usual assumption of a single factor.

(b) 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. (Cases d, g, and k.)

(c) If every item in a test is perfectly reliable the test is perfectly reliable even though all intercorrelations are equal to zero ($n$ equals $K$ and all $n_x$'s equal one) and regardless of the size of the sigmas (cases h, i, j, and k for the true $r_{tt}$) although the Kuder-Richardson would not indicate this fact, giving values all of the way down to zero for extreme cases.

(d) The size or uniformity of the sigmas is not important if the $n$'s and $r$'s are equal (Case e) since they then drop out of the formula. This indicates that they are the least important of the three determiners.

(e) The Brown-Spearman formula underestimates the true reliability by the ratio of $(n-K)/(n-1)$. (Case g.) The derivation of this formula is simple and easily understood. The usual Brown-Spearman formula reads

| Equal | Unequal | $r_{tt}$* | $r_{ttKR14}$ | Case |
|---|---|---|---|---|
| $n_x$ | $\sigma$ , $r$ | $\dfrac{n \sum r_{xx} \sigma_x^2}{K \sum \sigma_x^2 + (n-K) \sum r_{xx} \sigma_x^2}$ | $\dfrac{n-K}{n - K \dfrac{\sum \sigma_x^2}{(\sum \sigma_x)^2}} \cdot r_{tt}$ | a |
| $\sigma$ | $n_x$ , $r$ | $\dfrac{\sum n_x^2 r_{xx}}{n + \sum n_x (n_x - 1) r_{xx}}$ | $\dfrac{n}{n-1}\left[ r_{tt} - \dfrac{\sum n_x r_{xx}}{n + \sum n_x (n_x - 1) r_{xx}} \right]$ | b |
| $r$ | $\sigma$ , $n_x$ | $\dfrac{r_{aa} \sum n_x^2 \sigma_x^2}{\sum n_x \sigma_x^2 + r_{aa} \sum n_x (n_x - 1) \sigma_x^2}$ | $\dfrac{1}{1 - \dfrac{\sum n_x \sigma_x^2}{(\sum n_x \sigma_x)^2}} \left[ r_{tt} - \dfrac{r_{aa} \sum n_x \sigma_x^2}{\sum n_x \sigma_x^2 + r_{aa} \sum n_x (n_x - 1) \sigma_x^2} \right]$ | c |
| $n_x$ , $\sigma$ | $r$ | $\dfrac{n \sum r_{xx}}{K^2 + (n-K) \sum r_{xx}}$ | $\dfrac{n-K}{n-1} r_{tt}$ | d |
| $n_x$ , $r$ | $\sigma$ | $\dfrac{n r_{aa}}{K + (n-K) r_{aa}}$ | $\dfrac{n-K}{n - K \dfrac{\sum \sigma_x^2}{(\sum \sigma_x)^2}} r_{tt}$ | e (6) |
| $\sigma$ , $r$ | $n_x$ | $\dfrac{r_{aa} \sum n_x^2}{n + r_{aa} \sum n_x (n_x - 1)}$ | $\dfrac{n}{n-1}\left[ r_{tt} - \dfrac{r_{aa} \sum n_x}{n + r_{aa} \sum n_x (n_x - 1)} \right]$ | f |

| Equal | Unequal | $r_{tt}$* | $r_{ttKR14}$ | Case |
|---|---|---|---|---|
| $n_x, r, \sigma$ | — | $\dfrac{n\, r_{aa}}{K + (n-K) r_{aa}}$ | $\dfrac{n-K}{n-1} r_{tt}$ | g |
| $r = 1.00$ | $n_x, \sigma$ | $\dfrac{\sum n_x^2 \sigma_x^2}{\sum n_x \sigma_x^2 + \sum n_x (n_x - 1) \sigma_x^2} = 1.00$ | $\dfrac{1}{1 - \dfrac{\sum n_x \sigma_x^2}{(\sum n_x \sigma_x)^2}} \left[ 1.00 - \dfrac{\sum n_x \sigma_x^2}{\sum n_x \sigma_x^2 + \sum n_x (n_x - 1) \sigma_x^2} \right]$ | h |
| $r = 1.00, n_x$ | $\sigma$ | $\dfrac{n}{1 + (n-1)} = 1.00$ | $\dfrac{n-K}{n - K \dfrac{\sum \sigma_x^2}{(\sum \sigma_x)^2}}$ [1.00] | i |
| $r = 1.00, \sigma$ | $n_x$ | $\dfrac{\sum n_x^2}{n + \sum n_x (n_x - 1)} = 1.00$ | $\dfrac{n}{n-1} \left[ 1.00 - \dfrac{n}{\sum n_x^2} \right]$ | j |
| $r = 1.00, \sigma, n_x$ | — | $\dfrac{n}{K + (n-K)} = 1.00$ | $\dfrac{n-K}{n-1}$ [1.00] | k |

* Note that when the $n_x$ values are equal $n_x = n/K$, where $n$ is the total number of items in the test.

$$r_{tt_{BS}} = \frac{n\, r_{xx}}{1 + (n-1) r_{xx}} \tag{7}$$

where $r_{xx}$ equals the average intercorrelation for all the items, but for any given item under the condition of equal $n$'s, $\sigma$'s, and $r$'s for each factor group the average intercorrelation would be

$$r_{xx} = \frac{(n_a - 1) r_{aa} + (n - n_a) 0}{n - 1}, \tag{8}$$

and since $n_a$ would equal $n/K$ we would have

$$r_{xx} = \frac{n-K}{K} \frac{r_{aa}}{n-1}, \tag{9}$$

and substitution of this value in equation (7) yields

$$r_{tt_{BS}} = \frac{n-K}{n-1} \frac{n\, r_{aa}}{K + (n-K) r_{aa}}. \tag{6, g}$$

While this limiting error of $n-K/n-1$ for the two internal consistency methods of estimating reliability [see conclusions (b) and (e)] becomes negligible when $n$ becomes very large, it is nevertheless true that in short tests or when the formulas are used to estimate the reliability of test batteries where the number of tests is usually small, the equations based on internal consistency (assumption of one factor) would lead to gross underestimation and serious theoretical difficulty.

As to the gross underestimation, we present two cases of 12-item tests or batteries with assumed intercorrelations within the factor groups (the reliability of each item) equal to .95 for one test and to .60 in the other. For each test we shall assume anywhere from one to twelve factors with from twelve to one items, respectively. The true and estimated values follow:

| | | $r_{aa} = .95$ | | $r_{aa} = .60$ | |
|---|---|---|---|---|---|
| $K$ | $n_x$ | $r_{tt}$ | $r_{ttKR14}$ | $r_{tt}$ | $r_{ttKR14}$ |
| 12 | 1 | .950 | .000 | .600 | .000 |
| 6 | 2 | .974 | .531 | .750 | .409 |
| 4 | 3 | .983 | .715 | .818 | .595 |
| 3 | 4 | .987 | .808 | .857 | .701 |
| 2 | 6 | .991 | .901 | .900 | .818 |
| 1 | 12 | .996 | .996 | .947 | .947 |

An example of attendant theoretical difficulty due to the erroneous use of the internal consistency theory of reliability is the supposed dilemma cited by Royer (14) for the problem of multiple correlation. Low intercorrelations are a prerequisite to achieving high validity, whereas the consistency hypothesis holds that low intercorrelations mean low reliability. This of course does not follow at all if we make the probably nearer correct assumption of multiple orthogonal factors in both the criterion and the test items. It is of course true, as Thompson (16) has shown, using the correct reliability formula, that maximum validity and maximum reliability are not identical goals. If the criterion happens to have its variance determined in large part by factors predominating among the less reliable portions of the possible predictor items, then maximum validity will mean greatly lowered reliability as compared with the prediction of a different criterion whose factor composition is proportional to the reliabilities of the selected items, i.e., with a comparable test. We cannot follow Thompson's idea of compromising the weights in this case, however, since getting a better measure of something a person does not want to measure—a more comparable test—is no advantage if one thereby secures a poorer measure of what he is trying to measure—the criterion. The proper mode of approach in the Thompson situation would be to improve the reliability, by lengthening or other means, of those portions of the test battery with high regression weights, thereby increasing validity as well as reliability, rather than the Thompson suggestion of shifting the weights so as to secure higher reliability at the expense of decreased validity.

One other serious consequence of the erroneous adoption of the single factor assumption is a widespread misuse of the item-test correlation coefficient as a method of item selection and elimination. It is true that this measure, as Richardson (13) has clearly demonstrated, is the appropriate measure to use when there is a single factor among the items, but let us examine the values of this coefficient for various items under the condition of multiple orthogonal factors assumed above. In general this coefficient for any given item will equal

$$r_{it} = \sigma_i[1 - (n_i - 1)r_{ii}]/\sqrt{\sum n_x \sigma_x^2 + \sum n_x(n_x - 1)\sigma_x^2 r_{xx}}, \qquad (10)$$

where the subscript $i$ stands for the particular group of items having the same factor as the item in question. We see then that the item-test coefficient depends not on reliability ($r_{ii}$) alone but also upon $\sigma_i$ (item difficulty) and upon the number of items in the test measuring the same factor ($n_i$). Assuming these various determiners to be equal from group to group gives a series of values paralleling the

various conditions in set (6) above:

| Equal | Unequal | $r_{it}$ | Case |
|---|---|---|---|
| $n$ | $\sigma$ , $r$ | $\dfrac{\sigma_i[K + (n-K)r_{ii}]}{\sqrt{n}\sqrt{K\sum\sigma_x^2 + (n-K)\sum\sigma_x^2 r_{xx}}}$ | a |
| $\sigma$ | $n$ , $r$ | $\dfrac{1 + (n_i - 1)r_{ii}}{\sqrt{n + \sum n_x(n_x - 1)r_{xx}}}$ | b |
| $r$ | $\sigma$ , $n$ | $\dfrac{\sigma_i[1 + (n_i - 1)\bar{r}_{ii}]}{\sqrt{\sum n_x\sigma_x^2 + \bar{r}_{ii}\sum n_x(n_x - 1)\sigma_x^2}}$ | c (11) |
| $n$ , $\sigma$ | $r$ | $\dfrac{K + (n-K)r_{ii}}{\sqrt{n}\sqrt{K^2 + (n-K)\sum r_{xx}}}$ | d |
| $n$ , $r$ | $\sigma$ | $\sigma_i\sqrt{\dfrac{K + (n-K)r_{ii}}{n\sum\sigma_x^2}}$ | e |
| $\sigma$ , $r$ | $n$ | $\dfrac{1 + (n_i - 1)\bar{r}_{ii}}{\sqrt{(1 - \bar{r}_{ii})n + \sum n_x^2}}$ | f |
| $n$ , $r$ , $\sigma$ | — | $\sqrt{\dfrac{K + (n-K)\bar{r}_{ii}}{Kn}}$ | g |
| $r = 1.00$ | $n$ , $\sigma$ | $\dfrac{n_i\sigma_i}{\sqrt{\sum n_x^2\sigma_x^2}}$ | h |
| $r = 1.00$ , $n$ | $\sigma$ | $\sigma_i\sqrt{1/\sum\sigma_x^2}$ | i |
| $r = 1.00$ , $\sigma$ | $n$ | $n_i/\sqrt{\sum n_x^2}$ | j |
| $r = 1.00$ , $\sigma$ , $n$ | — | $\sqrt{1/K}$ | k |

From these equations in set (11) we can draw the following conclusions:

(a) The item-test coefficient is not a measure of item reliability alone, but depends upon the share of the total variance of the battery determined by the sub-battery of which it is a part.

(b) Perfectly reliable items with low $n$'s and $\sigma$'s would be dis-

carded as worthless by the usual method of applying this criterion, while less reliable items with large $n$'s and $\sigma$'s would be retained. Items with the highest $r_{it}$ value are not necessarily the best items.

(c) The value of $r_{it}$ does not approach 1.00 as an upper limit as the items become more reliable, but instead, ($n$'s and $\sigma$'s being equal) approaches the value of $\sqrt{1/K}$. Of course if $K = 1$, $r_{it}$ does approach 1.00, but if $K$ is greater than one the possible upper limits are

| $K$ | $r_{ii} = 1.00$ Case k upper limit of $r_{tt}$ | $r_{ii} = .50$ Case g |
|---|---|---|
| 2 | .707 | .505 |
| 4 | .500 | .361 |
| 6 | .408 | .297 |
| 8 | .354 | .260 |
| 10 | .316 | .234 |
| 20 | .224 | .173 |

The automatic rejection of items with low $r_{it}$ values is not justified.

(d) If $r_{ii}$ is equal to zero, the lower limit of the value of $r_{it}$ approaches $\sqrt{1/n}$ rather than zero, which makes the value of the Kuder-Richardson formula (8) (their article) fictitiously high. According to their concept of the single factor their formula (8) should, as their formula (14) does, become equal to zero when all of the intercorrelations are equal to zero, but this is not the case. Their formula (8) is

$$r_{tt_{KR8}} = \frac{\sigma_t^2 - \sum \sigma_i^2}{2\sigma_t^2} + \sqrt{\frac{\sum r_{it}^2 \sigma_i^2}{\sigma_t^2} + \left(\frac{\sigma_t^2 - \sum \sigma_i^2}{2\sigma_t^2}\right)^2}, \qquad (12)$$

which if we assume one factor with all $\sigma$'s equal becomes

$$r_{tt_{KR8}} = \frac{\sigma_t^2 - n\sigma_i^2}{2\sigma_t^2} + \sqrt{\frac{\sigma_i^2 \sum r_{it}^2}{\sigma_t^2} + \left(\frac{\sigma_t^2 - n\sigma_i^2}{2\sigma_t^2}\right)^2}. \qquad (13)$$

If we let all $r_{xx}$ values be 1.00, we have

$$\sigma_t^2 = n^2 \sigma_i^2 \quad \text{and} \quad r_{it} = 1.00,$$

whence

$$r_{tt_{KR8}} = \frac{n^2\sigma_i^2 - n\sigma_i^2}{2n^2\sigma_i^2} + \sqrt{\frac{n\sigma_i^2}{n^2\sigma_i^2} + \left(\frac{n^2\sigma_i^2 - n\sigma_i^2}{2n^2\sigma_i^2}\right)^2} = \underline{1.00}, \tag{14}$$

which is the correct answer. But if we let all $r_{xx}$ values equal .00 we have

$$\sigma_t^2 = n\sigma_i^2 \quad \text{and} \quad r_{it} = 1/\sqrt{n}$$

whence

$$r_{tt_{KR8}} = \frac{n\sigma_i^2 - n\sigma_i^2}{2n\sigma_i^2} + \sqrt{\frac{\sigma_i^2}{n\sigma_i^2} + \left[\frac{n\sigma_i^2 - n\sigma_i^2}{2n\sigma_i^2}\right]^2} = \sqrt{1/n} > .00, \tag{15}$$

which is incorrect. While this is the upper limit of error in this formula, it does remain spuriously large for all values of $r_{xx}$ less than 1.00. The source of this spurious increment is the incorrect assumption of a self-correlation of 1.00 for $r_{ii}$ in the item-test coefficient. While this is correct for the actual correlation with the test of which it is a part, the substitution of this value in the Kuder-Richardson series assumes it to hold as well for the "comparable" test series, the capital letter test, whereas the correct value here is not 1.00 but the actual reliability of the item. Thus when all $r_{xx}$ values are unity the reliability is 1.00 and formula (8) is correct, but for all other values of $r_{xx}$ the reliability will be less than unity and formula (8) will be in error, this error reaching a maximum as $r_{xx}$ approaches zero.

Some numerical examples of typical solutions for 3-factor problems are presented here as an indication of the points made under conclusions (a), (b), and (c) above:

| Case | $n$ | $\sigma$ | $r$ | $r_{it}$ | Case | $n$ | $\sigma$ | $r$ | $r_{it}$ |
|---|---|---|---|---|---|---|---|---|---|
| | 50 | .3 | .90 | .308 | | 50 | .5 | .20 | .105 |
| a | 50 | .4 | .70 | .326 | d | 50 | .5 | .70 | .303 |
| | 50 | .5 | .20 | .141 | | 50 | .5 | .90 | .383 |
| | 100 | .4 | .20 | .362 | | 50 | .5 | .60 | .559 |
| b | 40 | .4 | .70 | .492 | e | 50 | .4 | .60 | .447 |
| | 10 | .4 | .90 | .158 | | 50 | .3 | .60 | .335 |
| | 100 | .5 | .60 | .362 | | 100 | .5 | .60 | .534 |
| c | 40 | .4 | .60 | .492 | f | 40 | .5 | .60 | .216 |
| | 10 | .3 | .60 | .158 | | 10 | .5 | .60 | .057 |

| Case | $n$ | $\sigma$ | $r$ | $r_{it}$ | Case | $n$ | $\sigma$ | $r$ | $r_{it}$ |
|---|---|---|---|---|---|---|---|---|---|
| | 100 | .5 | 1.00 | .940 | | 100 | .4 | 1.00 | .887 |
| h | 40 | .4 | 1.00 | .304 | j | 40 | .4 | 1.00 | .355 |
| | 10 | .3 | 1.00 | .057 | | 10 | .4 | 1.00 | .089 |
| | 50 | .5 | 1.00 | .421 | | | | | |
| i | 50 | .4 | 1.00 | .337 | | | | | |
| | 50 | .3 | 1.00 | .253 | | | | | |

In view of the general possible unsatisfactory condition resulting from the application of the present methods of estimating reliability of the total test and of item validation by the blind assumption of a single factor, it seems advisable to suggest a marked revision in the present methods of test analysis. The obvious solution would be the factorial analysis of each test, but while ideal this would be very laborious and often impracticable, especially if the number of items were at all large. The calculation of the intercorrelations would alone be a tremendous undertaking to say nothing of securing the residual matrices and the final rotation of the obtained loadings. Let us consider simpler possibilities.

A start toward such a method is found in the concept of "item synonymization" advanced by Lentz and Whitmer (10). In their method item intercorrelations have to be computed and there are no clear-cut standards for inclusion of an item in any given synonymy. However, they have demonstrated that

(a) test items do tend to fall into groups,

(b) an item correlates more highly with its synonymy than with other synonymies, and

(c) synonymies tend to correlate lowly with each other.

If these synonymies were at all numerous in a test a much better estimate of its over-all reliability could be obtained by considering the synonymies as sub-tests, computing their reliabilities by the Kuder-Richardson formula( justified for such a consistent group), and then computing the total test reliability by using these coefficients together with the inter-synonymy correlations in their general formula (2). Item validity would be evaluated not in terms of the correlation with the total test put in terms of the item correlation with *any* of the synonymies.

The present writers suggest tentatively an approach based on item-test rather than inter-item correlation coefficients. The following steps constitute a job-analysis of the proposed method:

(1) Compute the total test score ($T$) using unit weights.

(2) Compute all item-test correlations ($r_{iT}$).

(3) Skim off the items with the highest such coefficients.

(4) Rescore the papers on the basis of the selected items ($S_I$).

(5) Compute all item correlations with the new score ($r_{iS_I}$).

(6) Add new items where $r_{iS_I} > r_{iT}$ and drop items where $r_{iS_I} < r_{iT}$ to form $S_I'$.

(7) Rescore the papers on the basis of the amended list ($S_I'$).

(8) Repeat steps (5) and (6), computing $r_{iS_I}'$, adding items where $r_{iS_I}' > r_{iS_I}$ and dropping them when $r_{iS_I}' < r_{iS_I}$.

(9) Compute $S_I''$ and repeat steps (7), (8), and (9) until no further changes are indicated, computing $S_I'''$, $S_I''''$, $\cdots$, $S_I^m$.

(10) Record the final $r_{iS_I}^m$ values and compute a new residual score $T' = T - S_I^m$.

(11) Repeat steps (2) through (9) using $T'$ in place of $T$ and $S_{II}^X$ in place of $S_I^X$.

(12) Record the final $r_{iS_{II}}^m$ values and compute a new residual test score $T'' = T' - S_{II}^m$ and continue as before.

(13) Repeat the entire process through $T^m$ until the test items are exhausted or until all $r_{iT}^m$ values approximate zero.

The selected sub-tests $S_I^m$, $S_{II}^m$, $S_{III}^m$, etc., will correspond to the item synonymies of Lentz, will tend to lie along pre-rotated orthogonal axes (if the $i$ items possessed simple structure to begin with), and the $r_{iS_x}^m$ values will be the factor loadings of the individual items on those axes.

We can illustrate this method of taking the three-factor numerical examples given for case (f) in the discussion of $r_{it}$, above. Here the correlations of the items with the total test would be

100 values of .534,
40 values of .216,
and 10 values of .057.

We would select the 100 highest items as $S_1$, rescore the tests, and compute the $r_{iS_1}$ values, obtaining:

100 values of .602,*
and 50 values of .000,

completing that phase. We would next compute $T - S_1 = T'$ and compute $r_{iT'}$, obtaining

100 values of .000,
40 values of .589,
and 10 values of .155.

We would then form $S_2$ from the 40 items with correlations of .589, rescore the papers, and then compute $r_{iS_2}$, obtaining

110 values of .000,
and 40 values of .607,*

completing the second phase. We would next compute $T'' = T' - S_2$ and compute $r_{iT''}$, obtaining

140 values of .000,
10 values of .627.*

Using these 10 values to form $S_3$, since rescoring would result in identical values, would complete the analysis of the test

The Lentz technique would have arrived at identical sub-tests but at a cost of $150 \times 149/2 = 11{,}175$ correlation coefficients rather than the present $6 \times 150 = 900$ sub-calculations. Of course all tests would not possess simple structure nor would they all contain an hierarchical arrangement such that $n_{s_1} > n_{s_2} > n_{s_3}$, etc. In case the several $n$'s or the larger of these approached equality or in case the contribution to the variance approached equality the separation of the factors would become more laborious and perhaps impossible.

Mosier (11) has warned against the use of $r_{it}$ in a two-factor test when (1) the factors are numerically equal, and (2) the items do not possess simple structure, i.e., have loadings on both factors. His general thesis is in line with the main argument of this paper, which extends the consideration to more than two factors and applies the idea to the whole problem of reliability as well as to item analysis. That his warning is pertinent as a criticism or possible limitation to the method of test analysis proposed above is also recognized.

* Computing the correlation of the item with the total not counting the item in question would produce the true value of .60 here and in later synonymies, but such error will not obscure the major relationships.

In two attempts to use this method empirically the writers found one test which worked out quite smoothly due to unequally represented factors, whereas in a second test the two main factors were equipotent and came out as a single sub-test until $S_I$ was broken up into $S_{IA}$ and $S_{IB}$ by visual inspection, after which the method proceeded to work satisfactorily.

The present writers intend to examine further the application of the suggested method of analysis to both theoretical and empirical tests. If it can be made to overcome the difficulties involved in equipotent factors it may well provide a practicable means of approximate factorial analysis for test situations where the usual methods would be prohibitive. The results of such research will be published in a later paper.

An article by Kelley (8), published after the beginning of the present article, must be given special mention since it foreshadowed empirically a number of the rational conclusions presented above. His finding that formulas (14), (20), and (21) of the Kuder-Richardson series yielded, for a three-item test with inter-correlations of zero, values of .00 for the reliability coefficient, while their formula (8) yielded a value of .58, is a perfect example of our equation (15) showing the erroneous nature of their formula (8) and of the incorrectness of their general approach when the number of factors, $K$, is large in comparison with the number of items, $n$. Since the Kelley article did not point out the incorrectness of their formula (8) and especially since he repeated their claim that it was their most reliable equation, it was feared that his reporting of the value .58 might be taken as lending support to that spurious equation.

While we agree that the Kuder-Richardson series is a measure of coherence rather than of reliability, we cannot accept Kelley's coefficient of coherence, $VC/SVC$, since it gave a value of .33 for a test whose coherence is obviously zero. The reason for this erroneous result is the same as that responsible for the error in the Kuder-Richardson formula (8). They used unity for the item reliabilities while the Kelley method of factor analysis is equivalent to putting these same fictitious unities in the diagonals of the factor matrix. Had Kelley used the correct communalities of zero, i.e., had he used the Thurstone approach, he would have attained the correct coefficient of coherence of .00 as yielded by the Kuder-Richardson equations (14), (20), and (21). The fact that he obtained .33 while the Kuder-Richardson formula (8) yielded $\sqrt{.33}$ or .58 is empirical evidence of the similarity of error.

We also cannot agree to Kelley's appeal to an "act of judgment" on the part of the experimenter, as in splitting a test in half or other

smaller fractions, as a valid or dependable method of computing reliability. The suggestion of a complete factor analysis as a basis for setting up sub-tests or the short sub-test selection method schematized above (if it proves practicable) with the then proper use of the basic Kuder-Richardson equation appears to be a sounder and much more scientific approach.

## REFERENCES


1. Brown, W. Some experimental results in the correlation of mental abilities. *Brit. J. Psychol.*, 1909-1910, 3, 296-322.
2. Edgerton, H. A., and Kolbe, L. E. The method of minimum variation for the combination of criteria. *Psychometrika*, 1936, 1, 183-187.
3. Edgerton, H. A., and Thomson, K. F. Test scores examined with the lexis ratio. *Psychometrika*, 1942, 7, 281-288.
4. Horst, P. Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika*, 1936, 1, 53-60.
5. Hotelling, H. The most predictable criterion. *J. educ. Psychol.*, 1935, **26**, 139-142.
6. Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
7. Jackson, R. W. B. Reliability of mental tests. *Brit. J. Psychol.*, 1938-39, 29, 267-287.
8. Kelley, T. L. The reliability coefficient. *Psychometrika*, 1942, 7, 75-83.
9. Kuder, G. F., and Richardson, M. W. The theory of the estimation of reliability. *Psychometrika*, 1937, 2, 151-160.
10. Lentz, T. F., and Whitmer, E. F. Item synonymization: a method for determining the total meaning of pencil-paper reactions. *Psychometrika*, 1941, 6, 131-139.
11. Mosier, C. A note on item analysis and the criterion of internal consistency, *Psychometrika*, 1936, 1, 275-282.
12. Paulsen, G. B. A coefficient of trait variability. *Psychol. Bull.*, 1931, 28, 218.
13. Richardson, M. W. Note on the rationale of item analysis, *Psychometrika*, 1936, 1, 69-76.
14. Royer, E. B. Some recent developments in test construction. *Proc. Okla. Acad. Sci.*, 1936, 16, 107-109.
15. Spearman, C. Correlation calculated from faulty data. *Brit. J. Psychol.*, 1909-1910, 3, 271-295.
16. Thompson, G. A. Weighting for battery reliability and prediction. *Brit. J. Psychol.*, 1939-40, 30, 357-366.
17. Thouless, R. H. Test unreliability and functional fluctuation. *Brit. J. Psychol.*, 1935-36, 26, 325-343.
18. Woodrow, H. Quotidian variability. *Psychol. Rev.*, 1932, 32, 245-256.