NOTES ON THE RATIONALE OF ITEM ANALYSIS

M. W. RICHARDSON

Item Validity

There is increasing use of item analysis procedures for the improvement of objective examinations. The development of the procedures of item analysis has consisted chiefly of the invention of various forms of an index of association between the test item and the total test score. At least ten indices of item validity have appeared in various articles, which have been chiefly concerned with the relative effectiveness of the indices as devices for the improvement of tests. (4, 5, 6, 10). Since these indices of "item validity" are substitutes for or approximations to the ordinary coefficient of correlation between the item and the total test score, it may be useful to present certain deductions from simple correlational algebra. The present writer is of the opinion that the ingenuity displayed in the invention of new indices has outstripped the critical examination of the logical foundation for item analysis. The subsequent discussion is therefore concerned only with the underlying rationale of item analysis.

The first step in the description of item analysis procedures is to express the item-test coefficient in terms of the item intercorrelations. A test score t is defined by the equation

$$t \equiv x_1 + x_2 + x_3 + \cdots + x_n \ , \tag{1}$$

where t is the deviate score on the test, and the x's are the deviate scores on the items, which are n in number. This definition embodies, of course, the usual practice of summing the unit or zero scores on the separate objective items to obtain the total test score. Let us take r_{it}, the correlation between any item i and the test t as a measure of item validity.

Then

$$r_{it} \equiv \frac{\sum x_i t}{N \sigma_i \sigma_t} \tag{2}$$

where σ_i is the standard deviation of item i, and σ_t is the standard deviation of the test scores. The general subscript i means that the formula applies to any item of a given test. The summation is over the population N. Substituting in (2) the value of t from (1), we have

$$r_{it} = \frac{\sum x_i (x_1 + x_2 + \cdots + x_n)}{N\sigma_i\sigma_t} \tag{3}$$

$$= \frac{\sum x_i x_1 + \sum x_i x_2 + \cdots + \sum x_i x_n}{N\sigma_i\sigma_t}$$

$$= \frac{N r_{i1}\sigma_i\sigma_1 + N r_{i2}\sigma_i\sigma_2 + \cdots + N r_{in}\sigma_i\sigma_n}{N\sigma_i\sigma_t}$$

If we now assume that

$$\sigma_i = \sigma_1 = \sigma_2 = \sigma_3 = \cdots \sigma_n \ ,$$

which is rigidly true when all items are of the same difficulty as measured by the percentage of correct response, and approximately true for a wide range of difficulty, we have

$$r_{it} = \frac{N\sigma_i^2 \sum_{k=1}^{n} r_{ik}}{N\sigma_i\sigma_t} = \frac{\sigma_i \sum_{k=1}^{n} r_{ik}}{\sigma_t} \ , \tag{4}$$

in which the summation is over the n correlations of item i with each of the n items in turn. In order to further simplify equation (4), the standard deviation of test scores σ_t will be expressed in terms of the test elements. Squaring (1), and summing, we have

$$\sum t^2 = \sum x_1^2 + \sum x_2^2 + \cdots + \sum x_n^2 + 2\sum x_1 x_2 + \cdots + 2\sum x_1 x_n$$

$$+ \cdots + 2\sum x_{n-1} x_n = N\sigma_t^2 = N\sigma_i^2 \sum_{i=1}^{n} \sum_{k=1}^{n} r_{ik} \ . \tag{5}$$

The double summation indicates that all item intercorrelations are taken. We can simplify (5) to

$$\sigma_t = \sigma_i \sqrt{\sum_{i=1}^{n} \sum_{k=1}^{n} r_{ik}} \ , \tag{6}$$

Substituting this value of σ_t in equation (4), we may write

$$r_{it} = \frac{\sum_{k=1}^{n} r_{ik}}{+\sqrt{\sum_{i=1}^{n} \sum_{k=1}^{n} r_{ik}}} \ . \tag{7}$$

Equation (7) expresses any item-test correlation as a function of the item intercorrelations. As applied to any item i of a test homogeneous in difficulty, the item-test correlation is equal to the sum of the correlations of that item with all items of the test, divided by

the positive square root of the sum of all item intercorrelations. (In any actual test, the denominator of (7) will not be imaginary). Since the denominator is constant in any situation where item analysis procedures are employed, it can be concluded that:

In a test of uniform difficulty, the correlation of an item with the test is proportional to the average correlation of that item with each item of the test.

Since the item intercorrelation coefficients themselves form a distribution, it may be concluded that:

The rejection of items whose correlations with the test are relatively low raises the average intercorrelations of the remaining items.

The formal similarity of equation (7) to Thurstone's expression for the first factor loading for the Centroid Method is not accidental. (9). The first factor loading on the centroid is a measure of the correlation between a test and the sum or average of the tests in the battery. A similar interpretation may be made in the item analysis situation. The item-test coefficient measures the correlation between a variable (the item) and the sum or average of many such variables. In this context, the item-test coefficient is the "factor" loading of the item with an arbitrary test variable which is the sum of the items. These considerations make it possible to conclude that:

The item-test coefficient gives an indication of the extent to which the item measures what the test as a whole measures. The item-test coefficient merely tells whether or not an item is in step with other items of the test.

Item Validity and Test Reliability

If we assume, as in the foregoing, equal difficulty of items, the Spearman-Brown Formula might be used to estimate the reliability of a test of n items from $\bar{r}_{ik}$, the (average) correlation between two items. This is significant in connection with the effect of rejection of items with low item intercorrelations upon the reliability of the test. Let us take $\bar{r}_{ik}$, the average item intercorrelation as a measure of item reliability. Equation (7) gives the expression for any item-test coefficient. If we now add the n item-test coefficients we have

$$\sum_{i=1}^{n} r_{it} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{n} r_{ik}}{+\sqrt{\sum_{i=1}^{n}\sum_{k=1}^{n} r_{ik}}} = +\sqrt{\sum_{i=1}^{n}\sum_{k=1}^{n} r_{ik}} \quad . \tag{8}$$

The sum of the item-test coefficients is simply the positive square root of the sum of the item intercorrelations.

Writing equation (8) in terms of the respective average coefficients we have

$$n\bar{r}_{it} = +\sqrt{n^2\bar{r}_{ik}},$$

where $\bar{r}$ means the average of the respective r's.
This is simplified to

$$\bar{r}_{it} = +\sqrt{\bar{r}_{ik}},$$

or

$$\bar{r}_{it}^2 = \bar{r}_{ik}. \tag{9}$$

Substituting $\bar{r}_{ik}$ in the Spearman-Brown Formula, we may write

$$R = \frac{n\bar{r}_{ik}}{1 + (n-1)\bar{r}_{ik}}, \tag{10}$$

where R is the reliability coefficient. Solving for $\bar{r}_{ik}$, we have

$$\bar{r}_{ik} = \frac{R}{n - nR + R}. \tag{11}$$

Also, from (9) and (11), we may write

$$\bar{r}_{it} = \sqrt{\frac{R}{n - nR + R}}. \tag{12}$$

Equation (12) gives a direct solution for the mean item-test coefficient. Either the mean item-test coefficient or its square may be used as a measure of the cohesiveness or purity of the test. If equation (10) is used to compute the reliability coefficient from the mean item-test coefficient and the number of items, the estimate of the reliability coefficient will not be subject to the fluctuations in the value of R which are due to the arbitrary samplings of items to get the two split-halves. These fluctuations may be considerable in magnitude for different split-halves when the test is short. (2).

From the foregoing equations, it is possible to conclude that:

The rejection of items with low item-test correlations raises the reliability of a test, if the number of items is held constant. Whether the reliability coefficient will be raised absolutely, even with a reduced number of items, depends upon the dispersion of the original item intercorrelations. If this dispersion is great, extending to a num-

ber of negative values, it is theoretically possible to attain a higher reliability with a smaller number of items.

True Variance and Item Intercorrelation

An alternative way of expressing the relationship of item intercorrelation to reliability is here given for its illustrative value. The true variance can be expressed in terms of the number of items, their common standard deviation, and the average item intercorrelation. The test variance may be written:

$$\sigma_t^2 = n\sigma_i^2 + n(n-1)\,\bar{r}_{ik}\,\sigma_i^2 \ ,$$

which may be simplified to

$$\sigma_t^2 = n\sigma_i^2[1 + (n-1)\,\bar{r}_{ik}] \ . \tag{13}$$

Equation (13) is simply another way of writing equation (5). Since the true variance is given by

$$\sigma_\infty^2 = R\sigma_t^2 \ , \tag{14}$$

we obtain by substituting in equation (14) the estimates of R and σ_t^2 from (10) and (13) respectively,

$$\sigma_\infty^2 = \frac{n\bar{r}_{ik}}{1 + (n-1)\,\bar{r}_{ik}} \cdot n\sigma_i^2 \left[1 + (n-1)\,\bar{r}_{ik}\right] .$$

This can be simplified to

$$\sigma_\infty^2 = n^2\sigma_i^2\bar{r}_{ik} \ . \tag{15}$$

The conclusion is that:

For tests of homogeneous difficulty and constant length, the true variance is proportional to the average item intercorrelation.

Empirical Verification

It is hardly necessary to verify equation (7), since the verification must consist essentially of numerical substitution into each of two cognate algebraic formulas. Nevertheless, the following data are presented. Twenty-five objective items were selected from a long achievement test, in a completely random manner, except that they were of approximately the same difficulty. Table I gives the difficulty distribution of the items.

The mean score of the 100 subjects on the 25 item test was 9.36; the standard deviation was 4.24. The item-test correlations were computed by use of the formula for the point bi-serial coefficient (the

Pearson r) $r = \frac{M_p - M_q}{\sigma} \cdot \sqrt{pq}$, where M_p is the mean score of those passing the item, M_q is the mean score of those failing the item, σ is the standard deviation of the distribution of scores, p and q are the percentage of correct and incorrect answers, respectively. (7).

TABLE I

Percentage of correct answers	Number of items
35	1
36	6
37	6
38	6
39	5
40	1

The item intercorrelations were computed according to the formula

$$r = \frac{p_{12} - p_1 p_2}{\sqrt{p_1 q_1 p_2 q_2}},$$

where $p_1 =$ the percentage of population who give correct response on the first item,

$p_2 =$ the percentage of correct response on the second item,

$q_1 = 1 - p_1$,

$q_2 = 1 - p_2$,

$p_{12} =$ percentage of the population who give the correct response to both items.

Table II displays in parallel columns the two independently computed values of the item-test coefficients of correlation.

Summary

The foregoing development indicates that the reliability of a test may be improved by the use of the procedures of item analysis. Furthermore, such procedures will tend to make the test more pure or homogeneous, in the sense of conserving those items which have the largest intercorrelations. This is the only sense in which it may be said that the conserved items are more "valid" than the rejected items. (8, 10).

The use of item analysis procedures of the type described does not necessarily select items whose sums will give the best prediction

TABLE II

Item Number	Item-test Correlation	
	Computed by the formula $r = \frac{M_p - M_q}{\sigma} \cdot \sqrt{pq}$	Computed by equation (7)
1	.424	.424
2	.510	.517
3	.289	.287
4	.373	.376
5	.157	.181
6	.370	.385
7	.285	.285
8	.144	.148
9	.262	.254
10	.189	.202
11	.456	.454
12	.080	.079
13	.564	.561
14	.328	.326
15	.436	.438
16	.515	.514
17	.214	.218
18	.388	.387
19	.416	.421
20	.280	.284
21	.412	.410
22	.481	.477
23	.312	.309
24	.274	.272
25	.559	.556

Average discrepancy = 1.31 per cent of first computed value.

of an external criterion; Horst's Method of Successive Residuals is a solution of this problem. (3).

The University of Chicago,
Chicago, Illinois.

BIBLIOGRAPHY

1. ANDERSON, JOHN E., The Effect of Item Analysis upon the Discriminative Power of an Examination, *J. Appl. Psychol.*, 1935, 19, 237-244.
2. BROWNELL, WM. A., On the Accuracy with which Reliability May Be Measured by Correlating Test Halves, *J. Exper. Educ.*, 1933, 1, 204-215.
3. HORST, PAUL, Item Analysis by the Method of Successive Residuals, *J. Exper. Educ.*, 1934, 2, 254-263.
4. LENTZ, T. F., HIRSHSTEIN, BERTHA, and FINCH, J. H., Evaluation of Methods of Evaluating Test Items, *J. Educ. Psychol.*, 1932, 23, 344-350.
5. LINDQUIST, E. F., and COOK, W. W., Experimental Procedures in Test Evaluation, *J. Exper. Educ.*, 1933, 1, 163-185.
6. LONG, JOHN A., and SANDIFORD, PETER, and others: The Validation of Test Items, *Bull. Dept. Educ. Res. Ontario Coll. Educ.*, 1935, No. 3. pp. 126.
7. RICHARDSON, M. W., and STALNAKER, J. M., A Note on the Use of Bi-serial r in Test Research, *J. Gen. Psychol.*, 1933, 8, 463-465.
8. SMITH, MAX, The Relationship between Item Validity and Test Validity, *Teach. Coll. Contrib. Educ.*, 1934, No. 621. pp. vii + 40.
9. THURSTONE, L. L., The Vectors of Mind, Chicago: Univ. Chicago Press, 1935. pp. xv + 266.
10. ZUBIN, J., The Method of Internal Consistency for Selecting Test Items, *J. Educ. Psychol.*, 1934, 25, 345-356.