THE METHOD OF INTERNAL CONSISTENCY
FOR SELECTING TEST ITEMS

JOSEPH ZUBIN

Department of Psychology, New York State Psychiatric Institute and Hospital

In the validation of psychoneurotic inventories and other tests of the questionnaire type it is often impossible to obtain independent external criteria for item analysis. In such cases it is customary to devise a scoring key based upon logical considerations and to give a credit of one unit for each item answered "correctly" or in accordance with the logical key. After computing the total score of each individual, the items are validated against the total score as a criterion. If no association exists between the "correct" response and the criterion, the item is discarded since it adds nothing to the measuring quality of the test. If the item is found to be associated with the criterion, the degree of the relationship is investigated. In order to determine the degree of association some assumption must be made with regard to the distribution of the responses to the item, and if no such assumption is warranted, the degree of relationship cannot be established.

There are three techniques for carrying out this analysis—the bi-serial $r$, critical ratio, and association methods. Each of these methods of analysis is serviceable in determining the value of an item for a given test. The purpose of this paper is to compare the relative merits of these three methods.

There is one difficulty which is inherent in the Criterion of Internal Consistency and is common to all the methods. It results from the fact that the criterion itself includes the item under analysis. This brings about a spurious relationship between item and criterion. The method of removing this spuriousness will be indicated for each method.

---

1 The author is grateful to Professor Truman L. Kelley for reading parts of the manuscript and to Dr. E. E. Cureton and Dr. Jack W. Dunlap for offering many valuable suggestions.

2 The size of the index of association, whether it be $\chi^2$, the critical ratio, or some other index, is no indication of the degree of association. It merely indicates how certain the existence of the association is.
CRITICAL RATIO METHOD\(^4\)

This method consists of a comparison between the mean in total score of those who answered the item "correctly" and the mean of those who failed to answer it "correctly." If the mean of the "successes" is significantly larger than that of the "failures," the item is retained. If the difference is not reliable, the item is excluded, and if the difference is reliably negative, the item itself or the manner of its scoring is revised so as to eliminate the negative relationship with the total score. Since the item in question is included in the total score, the mean of the "successes" is necessarily greater than it would have been if the item were not included. This spurious increase in mean may cause a reliable difference to appear when in reality no such difference exists.

Let the subscript \( q \) represent the statistics of the "successes" and the subscript \( p \), the statistics of the "failures." Then the critical ratio for the difference between \( M_q \) and \( M_p \) is

\[
C = \frac{M_q - M_p}{\sqrt{\sigma^2_{Mq} + \sigma^2_{Mp}}}
\]

Letting a prime represent the statistics when the item under analysis, \( y \), is excluded from the total score

\[
C' = \frac{M_q' - M_p'}{\sqrt{\sigma^2_{Mq'} + \sigma^2_{Mp'}}}
\]

The removal of item \( y \) from the total score will affect the total score of the "successes" only, since only the "successes" were credited with it.

Hence

\[
M_q' = M_q - 1 \quad \text{and} \quad M_p' = M_p
\]

The standard deviation of the "failures" \( \sigma_p \), remains unchanged, since the scores of the "failures" remain unchanged. The standard

\(^4\) A complete discussion of this method will be presented in a forthcoming paper: "The Correlation Coefficient versus the Critical Ratio as Methods of Expressing Relationship" by Joseph Zubin and J. B. Maller.

\(^4\) The correlation between \( M_q \) and \( M_p \) in random samples is equal to zero, as shown in "Standard Errors of Statistics of Frequency Distributions" by J. B. Maller and Joseph Zubin, to be published.
deviation of the "successes" also remains unchanged, since each "success" was increased by a constant, unity, and adding a constant to a series of scores does not alter their standard deviation.

Hence,

\[ \sigma_{M_p} = \sigma_{M_p'} \text{ and } \sigma_{M_q} = \sigma_{M_q'} \]

\[ C' = \frac{M_q - M_p - 1}{\sqrt{\sigma^2_{M_q} + \sigma^2_{M_p}}} = C - \frac{1}{\sqrt{\sigma^2_{M_q} + \sigma^2_{M_p}}} \]  

(1)

Even if \( C' \) were zero, a reliable difference could be obtained if the sigma of the difference were equal to .3 or less, which is not unusual. This spuriousness may be readily eliminated by reducing the difference in the means by unity.

**COMPUTATION OF THE CRITICAL RATIO**

The computation of the critical ratio may be facilitated in the following manner. Since the mean of the total scores, \( M_s \) and its sigma, \( \sigma_s \) is constant for all the items, only the mean and sigma in total score of the "successes" for each item need be computed. The mean and sigma in total scores of the "failures" can be obtained as follows:

\[ pM_p + qM_q = M_s. \]

Hence

\[ M_p = \frac{M_s - qM_q}{p}; \quad p + q = 1. \]

The difference between the means can then be obtained as follows:

\[ M_q - M_p = \frac{M_s - M_p}{p} \]  

(II)

The standard deviation of the "failures," \( \sigma_p \) can be obtained as follows:

\[ \sigma_p^2 = p[\sigma_p^2 + \hat{d}_p^2] + q[\sigma_q^2 + \hat{d}_q^2] \]

where

\[ \hat{d}_p = M_p - M_s \text{ and } \hat{d}_q = M_q - M_s \]

Hence

\[ \hat{d}_q - \hat{d}_p = M_q - M_s \text{ and } \hat{d}_p = \frac{q}{p} \hat{d}_q \]

Substituting
\[ \sigma_x^2 = p \sigma_p^2 + \frac{q}{p} d_x^2 + q[\sigma_e^2 + d_e^2] \]

and

\[ \sigma_s^2 = \frac{1}{p} \left[ \sigma_x^2 - q \sigma_e^2 - \left( q + \frac{q^2}{p} \right) d_x^2 \right] \]

\[ = \frac{1}{p} \left[ \sigma_s^2 - q \sigma_e^2 - \frac{q}{p} d_s^2 \right] \]

and the critical ratio when the item under analysis is excluded from the total score, is

\[ C' = \frac{(d_x - p) \sqrt{N}}{\sqrt{p \left[ \sigma_s^2 + \frac{p - q^2}{q} \sigma_e^2 \right] - q d_s^2}} \]  

(III)

Hence, in order to apply the critical ratio method to the validation of test items, we need to compute separately for each item only two measures: The mean and the standard deviation of the "successes." The critical ratio for an item can then be determined by substituting these values in equation (III). If the assumption of normality for the dichotomous variable (item) is warranted and if the regression is linear and homoscedastic,\(^8\) the critical ratio obtained in the above manner can be converted into a correlation coefficient,\(^2\) equivalent to the bi-serial.

**BI-SEYRAL CORRELATION METHOD**

A second method for determining the validity of an item is to compute only the **means** in total score of the "successes" and of the "failures" and then to determine the bi-serial correlation between the item and the total score by means of the formula

\[ r_{xy} = \frac{M_x - M_y}{\sigma_x} \cdot \frac{pq}{\tilde{\pi}} \]

where \( M_x \) and \( M_y \) are, as in the previous section, the means in total score (including the item \( y \) whose correlation with the total is sought).

\(^8\) A partial check on the homoscedasticity of the distribution may be applied by noting whether \( \sigma_x = \sigma_y \) within the limits of sampling errors (suggested by Dr. E. E. Curenton).
of the "successes" and the "failures," respectively, \( p = \) proportion of "failures," \( q = \) proportion of "successes," \( z \) is the ordinate at the point of dichotomy in \( y \) in a unit normal distribution, and \( \sigma_z \) is the standard deviation of total scores.\(^6\) Again letting the primed subscripts represent the statistics of the distribution not including the item under analysis, \( y, r_{x'y} = \frac{M_x' - M_y'}{\sigma_{x'}} \cdot \frac{pq}{z} \).

now

\[
M_x' - M_y' = M_x - M_y - 1
\]

as shown previously

\[
\therefore r_{x'y} = \frac{M_x - M_y}{\sigma_{x'}} \cdot \frac{pq}{z} - \frac{pq}{2\sigma_{x'}}
\]

Now

\[
\sigma_{x'} = \sigma_{(x-y)} = \sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y}
\]

Under the assumptions required for the bi-serial \( r \), the standard deviation of the item is unity, and since

\[
(M_x - M_y)\frac{pq}{z} = r_{xy}\sigma_y
\]

\[
r_{x'y} = r_{xy}\frac{\sigma_x}{\sigma_{x'}} - \frac{pq}{2\sigma_{x'}}
\]

\[
r_{x'y} = \frac{1}{\sqrt{\sigma_x^2 + 1 - 2r_{xy}\sigma_x\sigma_y}} \left\{ r_{xy}\sigma_x - \frac{pq}{z} \right\}
\]

(TIV)

Hence even when no relationship whatsoever exists between \( x' \) and \( y \) \( (r_{x'y} = 0) \), the correlation between the item and the total including the item, will not vanish, but will be equal to \( \frac{pq}{2\sigma_x} \). The maximum value that \( \frac{pq}{z} \) can have is about .6. Hence, even when no relationship exists between the item and the "true" criterion, the apparent correlation due to the inclusion of the item under analysis in the total score can be as high as \( \frac{.6}{\sigma_z} \).

For computation purposes \( r_{x'y} \) may be rewritten

\(^6\) The symbols are those utilized by Kelley in Statistical Method, Macmillan, 1924, pp. 350-361.
The bi-serial correlation is of value only when the assumption of normality in the dichotomous variable is warranted. If the assumption of normality is not made, and the individuals falling in each of the two parts of the dichotomy are regarded as having identical scores, (1 for the successes and 0 for the failures) then the correlation coefficient for this two point distribution may be written—

\[ r'_{xy} = \frac{M_x - M_p}{\sigma_x} \left( \frac{1 - 0}{\sqrt{pq}} \right) = \frac{M_x - M_p}{\sigma_x} \sqrt{pq} \]

This coefficient \( r' \) bears the same relationship to the bi-serial coefficient, \( r \), that Yule's four-fold \( r \) bears to the tetrachoric \( r \). If we correct the equation for the spuriousness involved in correlating an item with the total including the item,

\[ r'_{xy} = \frac{M_x - M_p - 1}{\sigma_x} \sqrt{pq} = \frac{r'_{x\bar{y}} - \sqrt{pq}}{\sqrt{\sigma_x^2 + pq - 2r'_{x\bar{y}}\sigma_x\sqrt{pq}}} \]

This can be arrived at more directly by noting that\(^7\)

\[ r'_{x'y} = r'_{(x-y)y} = \frac{r'_{x'y} - \sqrt{pq}}{\sigma_{x-y}} \]

where \( y \) is the score on the item under analysis.

For computation purposes,

\[ r'_{xy} = \frac{[M_x - M_p - p]\sqrt{q/p}}{\sqrt{\sigma_x^2 - 2[M_x - M_p - \frac{p}{2}]q}} \]

**COMPUTATION OF BI-serial \( r \)**

The following procedure will facilitate the computation of the bi-serial \( r' \)s:


\(^8\) Suggested by Mr. Max Smith.
Method of Internal Consistency

1. Divide the total score of each individual by $\sigma_s$ and call the result $t$ (sigma scores).
2. Add together the $t$ scores for all the successes to obtain
   \[ m_q = \frac{M_q}{\sigma_s} \]
   The bi-serial $r$ between an item and the total score including the item is
   \[ r_{sy} = \frac{m_t - m_s}{p \frac{p}{z} = \frac{m_q - m_s}{z}} \]

Let the values of $(m_q - m_s)$ be designated by $d$. We can prepare a table for $d$ and $q/z$ and thus enable the determination of $r$ with very little effort. Since the value of $q/z$ is fixed for any value of $q$, Table I gives the values of $q$ instead of $q/z$. The table is used as follows: If $q = .30$ and $d = .70$, look in the column $q = .30$ until the value of $d$ approximating .70 is found—and read off the value of $r$ on the left hand margin $-.60$. If the nearest tailed value of $d$ is not close enough, interpolation may be resorted to. After the value of $r$ is obtained, it can be readily corrected for the spuriousness discussed above by means of equation (IV).

**Table I.** Bi-Serial $r$ for Given Percentage of “Successes” and Given Standard Difference Between Mean of “Successes” and Total Mean

<table>
<thead>
<tr>
<th>$r$</th>
<th>0.001</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.234</td>
<td>0.290</td>
<td>0.175</td>
<td>0.158</td>
<td>0.140</td>
<td>0.127</td>
<td>0.116</td>
<td>0.106</td>
<td>0.098</td>
<td>0.085</td>
<td>0.079</td>
</tr>
<tr>
<td>0.2</td>
<td>0.673</td>
<td>0.412</td>
<td>0.351</td>
<td>0.310</td>
<td>0.280</td>
<td>0.242</td>
<td>0.218</td>
<td>0.187</td>
<td>0.152</td>
<td>0.127</td>
<td>0.105</td>
</tr>
<tr>
<td>0.3</td>
<td>1.010</td>
<td>0.619</td>
<td>0.539</td>
<td>0.466</td>
<td>0.420</td>
<td>0.381</td>
<td>0.347</td>
<td>0.317</td>
<td>0.290</td>
<td>0.268</td>
<td>0.248</td>
</tr>
<tr>
<td>0.4</td>
<td>1.348</td>
<td>0.820</td>
<td>0.709</td>
<td>0.621</td>
<td>0.555</td>
<td>0.506</td>
<td>0.465</td>
<td>0.433</td>
<td>0.400</td>
<td>0.371</td>
<td>0.345</td>
</tr>
<tr>
<td>0.5</td>
<td>1.684</td>
<td>1.021</td>
<td>0.875</td>
<td>0.772</td>
<td>0.680</td>
<td>0.625</td>
<td>0.579</td>
<td>0.546</td>
<td>0.521</td>
<td>0.492</td>
<td>0.469</td>
</tr>
<tr>
<td>0.6</td>
<td>2.020</td>
<td>1.327</td>
<td>1.083</td>
<td>0.932</td>
<td>0.839</td>
<td>0.767</td>
<td>0.707</td>
<td>0.654</td>
<td>0.608</td>
<td>0.569</td>
<td>0.537</td>
</tr>
<tr>
<td>0.7</td>
<td>2.356</td>
<td>1.449</td>
<td>1.226</td>
<td>1.056</td>
<td>0.955</td>
<td>0.875</td>
<td>0.817</td>
<td>0.761</td>
<td>0.710</td>
<td>0.664</td>
<td>0.628</td>
</tr>
<tr>
<td>0.8</td>
<td>2.690</td>
<td>1.608</td>
<td>1.396</td>
<td>1.243</td>
<td>1.164</td>
<td>1.080</td>
<td>1.018</td>
<td>0.960</td>
<td>0.906</td>
<td>0.859</td>
<td>0.821</td>
</tr>
<tr>
<td>0.9</td>
<td>3.010</td>
<td>1.764</td>
<td>1.570</td>
<td>1.429</td>
<td>1.352</td>
<td>1.271</td>
<td>1.201</td>
<td>1.135</td>
<td>1.079</td>
<td>1.035</td>
<td>0.996</td>
</tr>
<tr>
<td>1.0</td>
<td>3.320</td>
<td>1.927</td>
<td>1.720</td>
<td>1.594</td>
<td>1.529</td>
<td>1.457</td>
<td>1.387</td>
<td>1.321</td>
<td>1.260</td>
<td>1.208</td>
<td>0.999</td>
</tr>
</tbody>
</table>

**Association Method**

The association method consists of comparing the proportion of “successes” among the individuals whose total scores lie above the median with the proportion of “successes” lying below the median.
This may be done either by computing $\chi^2$ for the four-fold table given below or by calculating the critical ratio for the difference in proportion of "successes" lying above and below the median respectively.\footnote{The equivalence of these two methods is proved by Professor Harold Hotelling in a forthcoming publication.}

<table>
<thead>
<tr>
<th></th>
<th>&quot;Successes&quot;</th>
<th>&quot;Failures&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above median</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Below median</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

This comparison involves two difficulties. First, since success in the item raises the total score, there may be some individuals whose scores would drop below the median if the item under analysis were omitted from the total score. Secondly, the median of the total scores is not a good reference point for all the items. The selection of the median as a reference point makes the process of item analysis rather simple and easy, but it tends to obscure some relationships and to enhance or create others. If we select an easy item, one that was passed by 85 individuals out of a total of one hundred, we find, if the item is a good one, that the "successes" tend to have higher scores than the "failures." Letting five of the high scoring individuals fail by chance and five of the low scoring individuals succeed, we should find eighty "successes" among the high scorers and five among the low scorers. But not all the eighty high scorers can fall above the median. Some of these must fall below the median. Table A shows the four-fold table for this case.

<table>
<thead>
<tr>
<th></th>
<th>&quot;Successes&quot;</th>
<th>&quot;Failures&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above median</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>Below median</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>15</td>
</tr>
</tbody>
</table>

$\chi^2$ for this table is 1.9 and the critical ratio of the difference of the proportions of "successes" lying above and below the median, respectively, is 2.0. If instead of the median, the 85th percentile were taken as the point of reference, Table B would result.
\[ \chi^2 \] for this table is thirty-six and the critical ratio for the proportion of "successes" is 11.5. It is apparent that the median is not a good reference point for all the items. The percentile in total score corresponding to the percentage of "successes" is better. The difficulty that was pointed out above will hold true equally well of both easy and difficult items. For, by interchanging the columns and the rows of the four-fold table, the item can be changed from a difficult one into an easy one, but the \[ \chi^2 \] value and the critical ratio remain unaltered.

In some instances the selection of the median as a point of reference results in the creation of spurious relationship. This can be seen most readily when we consider a case in which the items are responded to by chance—as for example, if a list of nonsense syllables are to be matched with right or wrong.\(^{10}\) The distribution of items by percentage passing (distribution of difficulty) will be normal and the distribution of total scores will be similarly normal. For any given item, the proportion of "successes" among the high scorers (lying above the median in total score) may be equal to the proportion of successes among the low scorers (lying below the median). But this can not hold true for all the items. Some items will be found whose proportion of successes among high scorers will be higher than the proportion of successes among the low scorers. Otherwise, average total score of those lying below the median will be equal to the average total score of those lying above the median, which is impossible. This situation is aggravated even further when instead of considering the median in total score as the point of reference, comparisons are made between the upper and lower quartile or between the highest and lowest decile in total score. This initial selection tends to create a correlation between the score on the item and the total score of an individual. In order to remove this spuriousness, it is better to take as the point of reference either the percentile in total score corresponding to the percentage of "successes," the median score of all the

\(^{10}\) Suggested by Dr. J. B. Maller.
The Journal of Educational Psychology

"successes" in the item in question, or the point in total score corresponding to the threshold of correctness in the given item.\(^{11}\) After choosing the reference point, \(\chi^2\) can be computed for the four-fold table.

<table>
<thead>
<tr>
<th></th>
<th>&quot;Successes&quot;</th>
<th>&quot;Failures&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above reference point</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Below reference point</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>q</td>
<td>p</td>
</tr>
</tbody>
</table>

If the assumption of continuity in the dichotomous variable is not warranted, the four-fold \(r\) can be obtained, and if normality can be assumed, the tetrachoric \(r\) can be computed.\(^{12}\)

**RELATIVE EFFICIENCY OF THE THREE METHODS**

Before comparing the efficiency of the three statistics—Critical Ratio, Bi-serial \(r\), and \(\chi^2\), it should be pointed out that measures of relationship fall into two major divisions—measures of the probability of association and measures of the degree of association. Measures of the first class indicate only whether or not an item is at all related to the criterion. Measures of the degree of association indicate in addition the relative importance of the item and can be utilized as weights.

When \(r_b\), the bi-serial correlation coefficient, is used as a measure of probability of association, it must be divided by its standard error. The critical ratio of the difference between the mean of successes and the mean of failures, \(C_s\), is always larger than \(C_r\left( = \frac{r}{\sigma_r}\right)\) when \(\chi^2 > 1 - r^2\), \(x\) being the sigma deviate of the point of truncation of the

\(^{11}\) This method is borrowed from psycho-physics and is used in threshold determination. Line has applied it as follows: The total scores are ranked in order of size and the point at which the number passing the given item equals the number failing, is selected as the reference point. See Line, W.: "The Growth of Visual Perception in Children." *Br. J. Psy. Mon. Supp.*, 1931, No. 16. Another way of selecting this point is to obtain the median of the range in which passers and failures alternate.

\(^{12}\) The computation of the tetrachoric \(r\) is simplified by the Alice Lee Tables in Biometrika; and diagrams for computing the tetrachoric \(r\) are given by Choire, L. Saffir, M. and Thurstone, L. L.: "Computing Diagrams for the Tetrachoric Correlation Coefficient." University of Chicago Bookstore, 1933.
Method of Internal Consistency

dichotomous variable in a normal distribution. Hence $C_s$ is more efficient than $C_r$ under the above conditions.

It must be remembered that $r_v$ requires the assumption of normality in the distribution of the item. Since $C_s$ requires no such assumption it is to be preferred on this additional account also. Furthermore, if the assumption of normality is warranted, $C_s$ can be transmuted into $r_v$ by means of the equation

$$r_v^2 = \frac{1}{1/C_s^2 - 1/C_r^2 + 1}$$

where $C_1$ is the critical ratio of the difference between the means of the "successes" and "failures" in the unit normal distribution of the item and $C_2$ is the unit critical ratio of the means of the "successes" and "failures" in total score (equal to the regular critical ratio divided by $\sqrt{N}$). Tables giving the value of $r$ for each value of $C_1$ and $C_2$ have been provided.

The association method makes the fewest number of assumptions. It does not, however, provide a direct measure of degree of association although such a measure is available in the tetrachoric $r$. But both $\chi^2$ and the tetrachoric $r$ are perforce less efficient than the critical ratio and the bi-serial $r$ since the latter two methods take into consideration the gradations in one of the variables, while the $\chi^2$ and tetrachoric $r$ consider both variables as dichotomous.

**SUMMARY**

Three methods for internal validation of test items have been reviewed—the critical ratio, bi-serial, and association methods. The spuriousness due to the customary practice of including the item under analysis in the total score was pointed out and formulae for correction were provided. In general, the association method, in which the percentages of "passes" lying above the median in total score becomes the measure of validity, is the easiest to apply but it is subject to serious limitations. Its modified form in which a reference point must be found for each item is superior but much the most difficult of all to compute. Next in order of ease of computation are the critical ratio, and the bi-serial $r$. The bi-serial $r$ method is subject to the

---

13 See footnote 3.

14 The true distributions of $C_s$ and $C_r$ are not available, and the tacit assumption is made that when $N$ is large they approach normality.

15 See footnote 3.
assumption of normality in the dichotomous variable, but a coefficient not involving this assumption, \( r' \), can be used. The critical ratio method is subject to no assumptions except those underlying the derivation of the standard error of a mean and is, therefore, the most useful measure. If the assumptions of linearity and homoscedasticity are warranted, the critical ratio can be transmuted directly into the bi-serial correlation coefficient.