THE FACTORIAL INTERPRETATION OF TEST DIFFICULTY

GEORGE A. FERGUSON

DEPARTMENT OF EDUCATIONAL RESEARCH, UNIVERSITY OF TORONTO

This paper discusses the influence of test difficulty on the correlation between test items and between tests. The greater the difference in difficulty between two test items or between two tests the smaller the maximum correlation between them. In general, the greater the number of degrees of difficulty among the items in a test or among the tests in a battery, the higher the rank of the matrix of intercorrelations; that is, differences in difficulty are represented in the factorial configuration as additional factors. The suggestion is made that if all tests included in a battery are roughly homogeneous with respect to difficulty existing hierarchies will be more clearly defined and meaningful psychological interpretation of factors more readily attained.

The presumption underlying recent developments in the theory of test structure is that the greater a test's internal consistency the greater its efficacy as an instrument for the measurement of mental ability. By internal consistency is meant that every item should correlate as highly as possible with every other item and as highly as possible with the test as a whole. The more closely this condition is satisfied the more closely the test approximates to the measurement of a unit trait. A test may be regarded as measuring a unit trait in the ideal case when the matrix of inter-item correlations is of rank 1 and when no specific variance other than error variance is found in the factorial configuration describing the inter-item correlation matrix; that is, all inter-item correlations when corrected for attenuation are in the neighbourhood of unity. Scrutiny of formulas for the correlation of sums indicates immediately that the more closely a test approximates to the measurement of a unit trait the greater its variance and the greater its reliability.

If the criterion of internal consistency, as I understand it, is to be reasonably approximated, the items in a test must be homogeneous with respect to difficulty, the difficulty of an item being described by the proportion of persons in a clearly defined population who pass it. Furthermore, the items must be homogeneous with respect to content; that is, all the items in the test must be of the same type. Relevant to the foregoing considerations is the observation that although in a given test the conditions above may seem to be fairly well satisfied the test may not be a satisfactory measuring instrument for
some specified purpose, since if all the items are of equal difficulty high discrimination may be secured at one particular level of ability at the expense of discrimination at other levels of ability.

Now the essential problem with which this paper is concerned is that if the items in a test, or the tests in a battery, are homogeneous with respect to content but heterogenous with respect to difficulty, the matrix of item intercorrelations, or test intercorrelations, will have a rank greater than 1, and cannot, therefore, from the factorial point of view be regarded as measuring a unit trait; indeed, it would seem that the greater number of degrees of difficulty the higher the rank of the correlation matrix.

Consider firstly the influence of difficulty on the correlation between two test items. The variance of a single test item is given by

\[ s_i^2 = p_i q_i \]  

where \( p_i \) is the proportion of persons passing item \( i \), and \( q_i \) is the proportion of persons failing item \( i \). The correlation* between two dichotomously scored test items is given by

\[ r_{ij} = \frac{P_{ij} - P_i P_j}{s_i s_j} \]  

where \( P_{ij} \) is the proportion of persons passing both items \( i \) and \( j \). When \( p_i = p_j \), the correlation \( r_{ij} \) has a maximum value of unity. When, however, \( p_i > p_j \) the quantity \( P_{ij} \) has a maximum value equal to \( p_j \), and the maximum value of \( r_{ij} \) is given by

\[ \sqrt{\frac{P_i q_j}{p_i q_j}} \]  

To illustrate: if \( p_i = .70 \) and \( p_j = .30 \), the correlation between two such items can never exceed .4286. In general, we may state that the greater the difference in difficulty between two test items, the smaller the maximum correlation between them.

Consider now a hypothetical test of \( n \) items arranged in ascending order of difficulty. Let the difficulty of the items be \( p_1, p_2, \ldots, p_n \). Let us assume that the \( x \) persons passing the \( i \)th item are the \( x \) persons making the \( x \) highest scores on the whole test of \( n \) items. Under this specified condition \( p_1 > p_2 > p_3 > \cdots > p_n \), and \( p_{12} = p_2, p_{13} = \) as follows:

\[ p_2, \ldots, p_{(n-1)n} = pn. \] Therefore, the item covariance \( p_{ij} = p_i q_j = p_i q_i \), where \( p_i > p_j \). The matrix of inter-item covariances is then

* The correlation coefficient used to indicate the relation between two dichotomously scored test items is arbitrary. The formula given here is identical with the coefficient \( \phi \) commonly used in dealing with fourfold point surfaces.