THE JOURNAL OF EXPERIMENTAL EDUCATION
(Volume 35, Number 4, Summer 1967)

THE MAXIMUM RELIABILITY OF A MULTIPLE-CHOICE TEST AS A FUNCTION OF NUMBER OF ITEMS, NUMBER OF CHOICES, AND GROUP HETEROGENEITY

GRAHAM J. BURKHEIMER
DONALD W. ZIMMERMAN
RICHARD H. WILLIAMS
East Carolina College

IN PREVIOUS papers (7, 8) it has been shown that chance success due to guessing introduces an unavoidable source of error into multiple-choice test scores. This particular class of error is negatively correlated with true scores. The usual equations for test reliability and other intercorrelations among components of test scores depend upon the assumption that the correlations between true scores and error scores and between error scores and error scores on parallel forms of a test are zero. In previous papers (6, 8, 9, 10) more general equations for these intercorrelation terms, which do not depend upon the above assumptions, have been presented.

Because of the presence of chance success due to guessing the reliability of a multiple-choice test has a maximum value. In other words, if all sources of error other than chance success due to guessing were eliminated, the reliability of a test would remain at some value less than unity because of the unavoidable error due to guessing. The computer simulation method described previously (8) gave reliabilities for several kinds of tests, under the assumption that only error due to guessing is present. The purpose of this paper is to determine these values using analytic methods. An equation for the maximum reliability of a multiple-choice test, which involves only number of items, number of choices, and mean and variance of true scores (group heterogeneity) is derived.

Horst (2) derived equations indicating the maximum correlation between two different tests. Beginning with these, Roberts (5) derived equations for maximum reliability of a test. These results involve item difficulties and are based on assumptions concerning intercorrelations among items. The relation of number of alternative choices to test reliability has also been investigated by Carroll (1), Lord (3), and Plumlee (4). The present paper differs from these approaches to the problem in that it does not involve item difficulties, but considers only components of variance of test scores. It involves no assumptions about intercorrelations among items and holds for the case in which there is a negative correlation between true scores and error scores introduced by guessing. The result is relatively simple in form.

VARIANCE OF ERROR SCORES AND OF OBSERVED SCORES

When chance success due to guessing is the only source of error, the error scores for those true scores having a fixed value, $\underline{T}$, will approach a binomial distribution as the number of cases considered increases without limit. Therefore, we can write

$$[1] \qquad \overline{E}_T = \frac{\sum_{i=1}^{K_T} E_{Ti}}{K_T} = np.$$

where $\overline{E}_T$ is the mean of the error scores for the true scores having some fixed value, E_{Ti} is an error score for one of thse true scores, and K_T is the number of true scores having that particular value.

Here np indicates the mean of a binomial distribution.

Since n = N - T and p = 1/a (7) we have

$$[2] \quad \sum_{i=1}^{K_T} E_{Ti} = \frac{K_T N - K_T T}{a},$$

where N is the total number of items and a is the number of choices per item.

Summating, as the true score value varies from O to N, we can write

$$[3] \quad \sum_{T=O}^{N} \sum_{i=1}^{K_T} E_{Ti} = \sum_{T=O}^{N} \frac{K_T N - K_T T}{a},$$

or,

$$[4] \quad \Sigma E = \frac{KN - \Sigma T}{a} .$$

Equation [4] gives, in other words, the sum of the error scores in the entire distribution of test scores.

The variance of error scores corresponding to the true scores for a fixed value, T, is

$$[5] \quad s^2_{e_t} = \frac{\sum_{i=1}^{K_T} E^2_{Ti}}{K_T} - \frac{\left(\sum_{i=1}^{K_T} E_{Ti}\right)^2}{K^2_T}.$$

As K_T increases without limit this variance is also given by the binomial formula, npq, where q= (1 - p), or $(1 - \frac{1}{a})$. Therefore we can write

$$[6] \quad \frac{\sum_{i=1}^{K_T} E^2_{Ti}}{K_T} - \frac{\left(\sum_{i=1}^{K_T} E_{Ti}\right)^2}{K^2_T} = (N - T)\left(\frac{1}{a}\right)\left(1 - \frac{1}{a}\right).$$

Solving [6] for $\sum_{i=1}^{K_T} E^2_{Ti}$ gives

$$[7] \quad \sum_{i=1}^{K_T} E^2_{Ti} = \frac{a-1}{a^2}\left[K_T N - K_T T\right] + \frac{\left(\sum_{i=1}^{K_T} E_{Ti}\right)^2}{K_T} .$$

Substituting [2] in [7] gives

$$[8] \quad \sum_{i=1}^{K_T} E^2_{Ti} = \frac{a-1}{a^2}\left[K_T N - K_T T\right] + \frac{1}{a^2}\left[K_T N^2 - 2K_T NT + K_T T^2\right].$$

Summating, as the true score value varies from O to N, leads to the following result:

$$[9] \quad \sum_{T=O}^{N} \sum_{i=1}^{K_T} E^2_{Ti} = \sum_{T=O}^{N} \frac{a-1}{a^2}\left[K_T N - K_T T\right] + \sum_{T=O}^{N} \frac{1}{a^2}\left[K_T N^2 - 2K_T NT + K_T T^2\right].$$

or

$$[10] \quad \Sigma E^2 = \frac{a-1}{a^2}(KN - \Sigma T) + \frac{1}{a^2}(KN^2 - 2N\Sigma T + \Sigma T^2).$$

Total error variance is given by

$$[11] \quad s_e^2 = \frac{\Sigma E^2}{K} - \frac{(\Sigma E)^2}{K^2}.$$

Substituting [4] and [10] in [11] and reducing gives

$$[12] \quad s_e^2 = \frac{1}{a^2} \left(\frac{\Sigma T^2 - \frac{(\Sigma T)^2}{K}}{K} \right) + \frac{a-1}{a^2} (N - \overline{T}), \text{ which can also be written as}$$

$$[13] \quad s_e^2 = \frac{s_t^2}{a^2} + \frac{a-1}{a^2} (N - \overline{T}).$$

In a similar manner it can be shown that the variance of observed scores is given by the following equation:

$$[14] \quad s_o^2 = \frac{(a-1)^2 s_t^2}{a^2} + \frac{a-1}{a^2} (N - \overline{T}).$$

Equations [13] and [14], then, give the variance of error scores and observed scored under the assumption that chance success due to guessing is the only source of error. These variances are expressed as a function of number of items, number of choices, and mean and variance of true scores.

CORRELATION BETWEEN ERROR SCORES ON PARALLEL FORMS OF A TEST

An expression will now be derived for the correlation between error scores on parallel forms of a test. This correlation can be written as follows:

$$[15] \quad r_{ee} = \frac{\Sigma e_1 e_2}{K s_e^2}, \text{ or}$$

$$[16] \quad r_{ee} = \frac{\Sigma E_1 E_2 - \frac{\Sigma E_1 \Sigma E_2}{K}}{\Sigma E^2 - \frac{(\Sigma E)^2}{K}}.$$

Expressions for ΣE and ΣE^2 are given in [4] and [10]. An expression is needed, therefore, for $\Sigma E_1 E_2$ in order to determine r_{ee}. We begin by finding the sum of $E_1 E_2$ values for a fixed E_1 value and a fixed T value. In other words, we consider a joint distribution or error scores on parallel forms of a test for each $\underline{T}$ value. We can write

$$[17] \quad \sum_{j=1}^{K_{E_2}} E_1 E_{2j} = E_1 \sum_{j=1}^{K_{E_2}} E_{2j}, \quad \text{since } E_1 \text{ is fixed. Using [2] gives}$$

$$[18] \quad \sum_{j=1}^{K_{E_2}} E_1 E_{2j} = E_1 K_{E_2} \left(\frac{N-T}{a}\right). \quad \text{Since for a fixed value of } E_1,\ K_{E_1} = K_{E_2}, \text{ we have}$$

$$[19] \quad \sum_{j=1}^{K_{E_2}} E_1 E_{2j} = E_1 K_{E_1} \left(\frac{N-T}{a}\right).$$

Summating now over the E_1 values gives the following:

$$[20] \quad \sum_{E_1=O}^{N-T} \sum_{j=1}^{K_{E_2}} E_1 E_{2j} = \sum_{E_1=O}^{N-T} E_1 K_{E_1} \left(\frac{N-T}{a}\right)$$

, which can also be written as

$$[21] \quad \sum_{E_1=O}^{N-T} \sum_{j=1}^{K_{E_2}} E_1 E_{2j} = \left(\frac{N-T}{a}\right) \sum_{i=1}^{K_T} E_{1i},$$

where K_T indicates the total number of cases for the fixed value of T.

Again using the equation [2] we have

$$[22] \quad \sum_{E_1=O}^{N-T} \sum_{j=1}^{K_{E_2}} E_1 E_{2j} = K_T \left(\frac{N-T}{a}\right)^2, \text{ or}$$

$$[23] \quad \sum_{E_1=O}^{N-T} \sum_{j=1}^{K_{E_2}} E_1 E_{2j} = \frac{K_T N^2 - 2K_T NT + K_T T^2}{a^2}.$$

We now need only summate over the T values to obtain $\Sigma E_1 E_2$ for the entire distribution. Doing this, we obtain

$$[24] \quad \Sigma E_1 E_2 = \sum_{T=O}^{N} \sum_{E_1=O}^{N-T} \sum_{j=1}^{K_{E_2}} E_1 E_{2j} = \sum_{T=O}^{N} \frac{K_T N^2 - 2K_T NT + K_T T^2}{a^2}, \text{ or}$$

$$[25] \quad \Sigma E_1 E_2 = \frac{1}{a^2} (KN^2 - 2N\Sigma T + \Sigma T^2).$$

Substituting equation [4] and [25] in the numerator of [16] and simplyfing, gives

$$[26] \quad r_{ee} = \frac{\Sigma T^2 - \frac{(\Sigma T)^2}{K}}{a^2 \left(\Sigma E^2 - \frac{(\Sigma E)^2}{K}\right)}.$$

Dividing by K in both numerator and denominator leads to the following result:

$$[27] \quad r_{ee} = \frac{s_t^2}{a^2 s_e^2}.$$

Reliability is given by

$$[28] \quad r_{oo} = 1 - \frac{s_e^2}{s_o^2}(1 - r_{ee})$$

(Reference 8). Substituting [27] in [28] we have

$$[29] \quad r_{oo} = \frac{a^2 s_o^2 - a^2 s_e^2 + s_t^2}{a^2 s_o^2}.$$

Subtracting [13] from [14] gives

$$[30] \quad a^2 s_o^2 - a^2 s_e^2 = (a-1)^2 s_t^2 - s_t^2.$$

Substituting this result in [29] and simplifying, we have

$$[31] \quad r_{oo} = \frac{(a-1)^2 s_t^2}{a^2 s_o^2}.$$

This expression gives maximum reliability in terms of number of choices, variance of true scores, and variance of observed scores. Substituting the value for s_o^2 given by [14] leads to the following alternative result:

$$[32] \quad r_{oo} = \frac{(a-1)\, s_t^2}{(a-1)\, s_t^2 + N - \overline{T}}.$$

This equation, then, gives the maximum reliability of a multiple-choice test as a function of number of items, number of choices, variance of true scores, and mean of true scores. It indicates that maximum reliability depends on group heterogeneity as well as test length and number of choices.

Since $\overline{O} = \overline{T} + \overline{E}$ and, from [4], $\overline{E} = \frac{N - \overline{T}}{a}$, we can write

$$[33] \quad \overline{T} = \frac{a\overline{O} - N}{a - 1}.$$

Solving [14] for s_t^2, substituting the results, together with [33], in [31], and simplifying gives another expression for maximum reliability:

$$[34] \quad r_{oo} = 1 - \frac{N - \overline{O}}{a s_o^2}.$$

ALTERNATIVE EQUATIONS FOR CORRELATION BETWEEN ERROR SCORES ON PARALLEL FORMS

Substituting [13] in [27] and simplifying, we have

$$[35] \quad r_{ee} = \frac{s_t^2}{s_t^2 + (a-1)(N - \overline{T})},$$

which is similar in form to [32]. Equation [34] can be written in this form:

$$[36] \quad s_o^2 (1 - r_{oo}) = \frac{N - \overline{O}}{a}.$$

Equation [28] can be written as follows:

$$[37] \quad s_o^2 (1 - r_{oo}) = s_e^2 (1 - r_{ee}).$$

Substituting the right hand side of [37] in [36] and simplifying, we have

$$[38] \quad r_{ee} = 1 - \frac{N - \overline{O}}{a s_e^2}.$$

COMPUTER CHECKS

The equations presented above give the values of r_{oo} and r_{ee} which would be expected if chance success due to guessing were the only source of error in multiple-choice tests. The reliabilities of actual tests would be expected to be less than these values because of the presence of other sources of error. In addition, if reliability were determined from a finite number of ordered pairs of observed scores on parallel forms of a test, with only error due to guessing present, there would be sampling variability of the reliability coefficient. The binomial distribution of error scores assumed in derivation of the equations, in other words, would be only approximated for any finite number of true scores.

As the number of ordered pairs of scores on parallel forms increases without limit, however, the reliability coefficient would be expected to come closer and closer to the values given by the equations. In a previous paper (8) a method of determining the reliability coefficient by a computer simulation method was described. It was shown that for fairly large numbers of scores (samples of 100, 400, 700, and 1000) the estimates given by the method were stable. For example, for ten samples of 400 scores, the reliability of a 100-item, two-choice test was indicated as .89, .88, .89, .87, .90, .88, .89, .89, .89, and .87.

In Table 1 the reliabilities given by the computer simulation method are compared to the values given analytically by equations [31], [32], and [34] above. Also, the correlations between error scores on parallel forms given by the computer program are compared to the values given by equations [27], [35], and [38] above. In making these checks we begin with a distribution of true scores having a certain mean and a certain variance. The computer program then generates error scores which depend upon the magnitude of the true scores, as a model of guessing error, and these are added to the true scores to give observed scores. Repeating the procedure gives results comparable to observed scores on parallel forms of a test, when guessing is the only source of error. Finally, product-moment correlations between the two sets of observed scores give an indication of test reliability. Also, correlation between the two sets of error scores is found, as well as the means and variances of all distributions.

It can be seen from the table that the values given by the computer program correspond closely to the values predicted from the equations presented in this paper.

REFERENCES

1. Carroll, J. B., "The Effect of Difficulty and Chance Success on Correlations Between Items

TABLE 1

COMPARISON OF COMPUTER RESULTS WITH VALUES PREDICTED FROM EQUATIONS

	N=10 a=2	N=10 a=5	N=100 a=2	N=100 a=5
r_{oo}*	.44	.74	.89	.97
r_{oo}**	.46	.77	.89	.96
r_{oo}***	.44	.76	.89	.97
r_{oo}****	.42	.76	.89	.97
r_{ee}^{1}	.46	.17	.89	.65
r_{ee}^{11}	.44	.15	.88	.67
r_{ee}^{111}	.44	.17	.89	.66
r_{ee}^{1111}	.45	.23	.89	.65

* Value obtained from computer program
** Value obtained by substituting computer data in equation [31]
*** Value obtained by substituting computer data in equation [32]
**** Value obtained by substituting computer data in equation [34]
1 Value obtained from computer program
11 Value obtained by substituting computer data in equation [27]
111 Value obtained by substituting computer data in equation [35]
1111 Value obtained by substituting computer data in equation [38]

or Between Tests", Psychometrika, X (1945), pp, 1-19.

2. Horst, P., "The Maximum Expected Correlation Between Two Multiple-Choice Tests", Psychometrika, XIX (1954), pp. 291-296.
3. Lord, F. M., "Reliability of Multiple-Choice Tests as a Function of Number of Choices per Item", Journal of Educational Psychology, XXXV (1944), pp. 175-180.
4. Plumlee, L. B., "The Effect of Difficulty and Chance Success on Item-Test Correlation and on Test Reliability", Psychometrika, XVII (1952), pp. 69-86.
5. Roberts, A. O. H., "The Maximum Reliability of a Multiple-Choice Test", Psychologia Africana, IX (1962), pp. 286-293.
6. Williams, R. H., and Zimmerman, D. W., "Some Conjectures Concerning the Index of Reliability and Related Quantities When True Scores and Error Scores on Mental Tests are Not Independent", The Journal of Experimental Education, XXXV, No. 2 (Winter 1966), pp. 76-79.
7. Zimmerman, D. W., and Williams, R. H., "Effect of Chance Success Due to Guessing on Error of Measurement in Multiple-Choice Tests", Psychological Reports, XVI (1965), pp. 1193-1196.
8. Zimmerman, D. W., and Williams, R. H., "Chance Success Due to Guessing and Non-independence of True Scores and Error Scores in Multiple-Choice Tests: Computer Trials with Prepared Distributions", Psychological Reports, XVII (1965), pp. 159-165.
9. Zimmerman, D. W., and Williams, R. H., "Independence and Non-independence of True Scores and Error Scores in Mental Tests: Assumptions in the Definition of Parallel Forms", The Journal of Experimental Education, XXXV, No. 3 (Spring 1967), pp. 59-64.
10. Zimmerman, D. W., Williams, R. H., and Rehm, H. H., "Test Reliability When Error Scores Consist of Independent and Non-independent Components", The Journal of Experimental Education, XXXV, No. 1 (Fall 1966), pp. 74-78.