Psychological Reports, 1965, 16, 1193-1196.

EFFECT OF CHANCE SUCCESS DUE TO GUESSING ON ERROR OF MEASUREMENT IN MULTIPLE-CHOICE TESTS

DONALD W. ZIMMERMAN AND RICHARD H. WILLIAMS

East Carolina College

Summary.—Chance success due to guessing is treated as a component of the error variance of a multiple-choice test score. It is shown that for a test of given item structure the minimum standard error of measurement can be estimated by the formula $\sqrt{(N-X)/a}$, where N is the total number of items, X is the score, and a is the number of alternative choices per item. The significance of non-independence of true score and this component of error score on multiple-choice tests is discussed.

The reliability of a test is limited by a number of factors which taken together are said to constitute "error." For actual tests these factors and the relative contribution of each are unknown.

One contribution to the error variance of a multiple-choice test score, however, is apparent from examination of the test. This is the chance error due to the guessing inherent in this type of test. Suppose a test consists of N items with a alternative choices for each item. If a person had no knowledge of the subject matter but marked the answers to all items with the aid of a table of random numbers, a score of $(1/a)N$ correct answers could be expected. If the number of alternatives per item is small (in most multiple-choice tests 4 or 5), a substantial component of the total score may be accounted for by successful guessing.

More important, however, is the *variability* of the component of the total score attributable to successful guessing. It is error variance which limits the reliability of a test. If all persons obtained the same number correct by guessing, there would be no problem. A constant would be added to the score. Different persons, however, will receive different increments in score as a result of more-or-less successful guessing. In fact, the number of correct guesses will presumably follow a binomial distribution with mean np (where n is the number of items guessed and p is the probability of a correct guess) and variance npq (where q is $1 - p$).

In scoring tests a "correction formula," in which a fraction of the number of wrong answers is subtracted from the number right, is sometimes used (cf. Lord, 1963). The correction makes the scores of those persons who guess more comparable to the scores of those who, for one reason or another, do not guess. It should be noted, however, that the correction has no effect upon variability introduced by guessing. For some scores, in other words, the formula will undercorrect, for others it will overcorrect.

The item structure of a test (the number of items and the number of alterna-

tive choices per item) can thus be considered as contributing a component of error of measurement which is unavoidable. A simple formula can be derived by which this value can be estimated for any particular test.

The following symbols will be used.

N = total number of items on the test
X = observed score
T = true score
a = number of alternative choices per item
n = number of items on which guesses are made
$s_{e\,min}^2$ = minimum error variance (variance of distribution of number of items guessed correctly)
$s_{e\,min}$ = minimum standard error of measurement.

The error variance in which we are interested is the variance of the distribution of number of items which are guessed correctly. This will be given by the binomial formula, npq, where n is the number of items on which guesses are made, p is the probability of a successful guess, and q is $1-p$. For a multiple-choice test p will be $1/a$, where a is the number of alternative choices per item, and q will be $1-(1/a)$.

In order to use the binomial formula the number of items on which guesses are made must be estimated. First, the conventional "correction formula" for guessing can be used to estimate the true score.

$$T = X-[1/(a-1)](N-X) \quad . \qquad [1]$$

Or, as usually expressed, the true score is estimated by subtracting a fraction of the number of items "wrong" from the number "right." This fraction is one divided by one less than the number of alternatives per item (¼ of number wrong for a test with 5 alternatives, ⅓ for 4 alternatives, and so on).

The number of items on which guesses are made can then be found by subtracting this result from the total number of items on the test.

$$n = N-\{X-[1/(a-1)](N-X)\} \quad . \qquad [2]$$

Finally, this result is substituted in the binomial formula to give the variance of the distribution of number of items guessed correctly,

$$s_{e\,min}^2 = [N-\{X-[1/(a-1)](N-X)\}](1/a)[1-(1/a)] \quad . \qquad [3]$$

Simplifying, the following result is obtained:

$$s_{e\,min}^2 = (N-X)/a \quad . \qquad [4]$$

The square root gives the minimum standard error of measurement,

$$s_{e\,min} = \sqrt{(N-X)/a} \quad . \qquad [5]$$

The value obtained by the formula is a minimum in that, if all other sources of error were eliminated, a standard error of this value would still be contributed by the item structure of the test.

Derivation of the formula is based on assumptions which are only approximated in an actual situation. Failure of these assumptions to hold precisely would further increase the standard error of measurement. For example, a person taking a test may eliminate some of the alternatives for a given item because he has partial information. There is never a sharp distinction between "knowing the answer" and "guessing" (cf. Horst, 1933). In an actual test, therefore, the probability of a successful guess will be somewhat greater than $1/a$.

Also, as said previously, there are many other sources of error in addition to guessing. Results obtained using the formula given above, therefore, must be considered as lower limits. Actual standard errors will be greater than the calculated values. The formula may prove useful, however, in giving a rough idea of what can be expected from any particular type of item structure.

For example, consider a test of 100 items, with 4 alternatives per item, and a score of 50. Use of the formula shows a minimum standard error of measurement of 3.5. Or, as an extreme example, consider a "true-false" test of 10 items and a score of 5. This is a special case of a multiple-choice test, where a is 2. Calculation shows a minimum standard error of measurement of 1.6. In this case the standard error would be almost as large as the standard deviation of the true scores which would be expected. Here the formula confirms what one would suspect, that short "true-false" tests are quite unreliable.

An important feature of this minimum standard error of measurement is that it varies with true score. The higher the true score, the lower its value will be. The standard error of measurement as usually understood is a fixed value for a given test. The confidence interval for true score which is established is the same width for any observed score. This difference reflects the special characteristics of the class of error variance considered in the present paper.

TABLE 1

MINIMUM STANDARD ERROR OF MEASUREMENT FOR A MULTIPLE-CHOICE TEST WITH N ITEMS AND a ALTERNATIVE CHOICES PER ITEM

N/a	2	3	4	5	N/a	2	3	4	5
10	1.6	1.3	1.1	1.0	90	4.7	3.9	3.4	3.0
20	2.2	1.8	1.6	1.4	100	5.0	4.1	3.5	3.2
30	2.7	2.2	1.9	1.7	110	5.2	4.3	3.7	3.3
40	3.2	2.6	2.2	2.0	120	5.5	4.5	3.9	3.5
50	3.5	2.9	2.5	2.2	130	5.7	4.7	4.1	3.6
60	3.9	3.2	2.7	2.4	150	6.1	5.0	4.3	3.9
70	4.2	3.3	3.0	2.6	200	7.1	5.8	5.0	4.5
80	4.5	3.7	3.2	2.8	250	8.0	6.5	5.6	5.0

Table 1 shows the values of the minimum standard error for selected values of N and a. These include those which would most often occur in tests. In calculating these values it has been assumed that the score being considered is ½ the total number of items.

An implication of the above consideration concerns non-independence of true score and error score in multiple-choice tests. In test theory it has been assumed often that error score and true score are uncorrelated. For multiple-choice tests where there is chance success due to guessing this assumption cannot be made. Those persons with low true scores will guess on more items and thus receive relatively higher error scores. On the other

hand, those persons with high true scores will guess on fewer items and receive lower error scores. Therefore, there will be a negative correlation between true score and error score. As shown above, minimum standard error of measurement is a decreasing function of true score.

The extent to which non-independence of true score and error score is a serious problem for test theory is not certain. Possibly the inaccuracy introduced by neglecting this relationship is not large. A similar situation has been found to be true in the case of other statistical problems where the fit of the theoretical model to the actual situation is imperfect (Box, 1953; Norton, 1953). In the case of multiple-choice tests, however, the fact of non-independence is clear and its possible effect could be large.

REFERENCES

BOX, G. E. P. Non-normality and tests on variances. *Biometrika*, 1953, 40, 318-335.

HORST, A. P. The difficulty of a multiple choice test item. *J. educ. Psychol.*, 1933, 24, 229-232.

LORD, F. M. Formula scoring and validity. *Educ. psychol. Measmt*, 1963, 23, 663-672.

NORTON, D. W. An empirical investigation of some effects of non-normality and heterogeneity on the *F*-distribution. Unpublished Ph.D. thesis in Education, State Univer. of Iowa. Reported in E. F. Lindquist, *Design and analysis of experiments in psychology and education*. Boston: Houghton-Mifflin, 1953.

Accepted May 10, 1965.