DOCUMENT RESUME

ED 053 223 24 TM 000 838

AUTHOR	Darlington, Richard B.; Cieslak, Paul J.
TITLE	Estimating the Validity of Educational Tests. Final Report.
INSTITUTION	Cornell Univ., Ithaca, N.Y.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO	BR-9-B-118
PUB DATE	Jun 71
GRANT	OEG-2-7-00-004(509)
NOTE	17p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Criterion Referenced Tests, *Educational Testing, *Evaluation Techniques, Hypothesis Testing, Mental Tests, *Predictive Validity, *Test Construction, Testing Problems, *Test Validity

ABSTRACT

A new variant of the standard method for estimating the accuracy of educational tests is examined. It is found that the estimates produced by the new method are essentially unbiased and that the typical sizes of the errors of the estimates approach their theoretical lower limit as size increases, though they are still noticably above it for small and moderate sample sizes. (Author)

BR 9-B-118
PA 24
Approved 8/17/71
TM

Dr. John Sokol
Director, Educational Research
DHEW-OFFICE OF EDUCATION - RM 1018
Federal Building
26 Federal Plaza
New York, New York 10007

FINAL REPORT
Project No. 9B118
Grant No. OEG-2-7-00-004(509)

ESTIMATING THE VALIDITY OF EDUCATIONAL TESTS

Richard B. Darlington and Paul J. Cieslak
Cornell University
Ithaca, N. Y. 14850

June 1971

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

1

Final Report

Project No. 9B118
Grant No. OEG-2-7-00-004(509)

Estimating The Validity of Educational Tests

Richard B. Darlington and Paul J. Cieslak

Cornell University

Ithaca, N. Y. 14850

June 1971

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

Table of Contents

0. Summary

Nontechnical summary. A new variant of the standard method for estimating the accuracy of educational tests is examined. It is found that the estimates produced by the new method are essentially unbiased and that the typical sizes of the errors of the estimates approach their theoretical lower limit as size increases, though they are still noticeably above it for small and moderate sample sizes.

Technical summary. A new variant of the cross-validation method, for estimating the validity of empirically-constructed tests, is examined. Validity estimates made by the new method are compared to the known validities of the tests. It is found that the estimates produced by the new method are essentially unbiased and that the standard error of estimate is within about 25% of its theoretical lower limit for sample sizes of 168, and within about 60% of its theoretical lower limit for sample sizes of 84. The new method seems like a more useful way of using data than does the standard cross-validation design.

1. Introduction and background

1.1 Introduction

The purpose of the present project is to check the accuracy of a new method, invented by the principal investigator, for evaluating the validity, or predictive power, of a certain class of mental tests known as empirically-constructed tests.

1.2 What is an empirically-constructed test?

An empirically-constructed test is a test in which actual data from subjects is used in the process of constructing the test. Specifically, the test-constructor has a sample of people with known scores on the variable to be predicted by the test (the criterion variable), and known scores on a large group of items or subtests. In educational settings, common criterion variables (that is, variables which investigators often attempt to predict) are grades in individual courses, grade-point averages, number of years a student will stay in school, etc. Items or subtests used to make the predictions may be previous grades, IQ scores, items in interest inventories, etc. The investigator uses statistical procedures to select a subset of items or subtests which, according to the available data, accurately predict the criterion variable.

1.3 How is an empirically-constructed test usually evaluated?

1.31 What statistic is used?

The simplest and most generally used statistic for evaluating a test is the coefficient of correlation between the test and the criterion variable. Although other statistics are in some cases theoretically preferable (e.g. see Darlington and Stauffer, 1966), they are usually found to rank-order the values of tests in almost exactly the same order as the simpler correlation coefficient (Darlington, 1967). Henceforth, when we speak of the "validity" of a test, we shall mean the correlation between the test and the variable it is designed to predict.

1.32 What sample of subjects is used?

It is well known (Cronbach, 1960, p. 355, or almost any other elementary text on test theory) that an empirically-constructed test has a substantially higher validity in the sample of subjects used in its construction than in other samples of subjects, or in the population of subjects from which the test-construction sample was drawn. Thus it is necessary to check the validity of an

empirically-constructed test by measuring its validity in a second sample of subjects, drawn from the same population as was the sample used in constructing the test. This process of checking a test's validity in a second sample of subjects is known as cross-validation.

1.4 What is wrong with the usual cross-validation technique for evaluating an empirically-constructed test?

The most serious problem with the procedure described in Section 1.32 (i.e. the standard cross-validation procedure) is that it is wasteful of subjects. If a total sample of N subjects (with known scores on the test items and the criterion variable) is available, the standard cross-validation procedure demands that this sample be split into two smaller subsamples, one subsample to be used for test construction, the other subsample to be used for cross-validation. If the original sample is small (as it often is, for practical reasons), the subsamples are even smaller. Thus the standard technique is wasteful of the already-small number of available subjects, in that only part of the subjects are used for test-construction, and another part for cross-validation.

1.5 What techniques have been invented to overcome the problems described in Section 1.4, and what are the shortcomings of those techniques?

1.51 The Wherry "validity shrinkage" formula

Most behavioral scientists have been taught to be extremely wary of any alternative to the traditional cross-validation technique for estimating the true validity of a test. Perhaps the major reason for this is that one of the earliest proposed alternatives to cross-validation was based on an error. This proposed alternative was the Wherry formula (1931). Although the error was not widely known, many writers (e.g. Guilford, 1954, p. 405)
observed that the Wherry formula didn't seem to work when its predictions were checked against actual observations. This observed discrepancy has tended to enshrine the rule "Don't try any alternatives to traditional cross-validation". Several other workers have noticed Wherry's mistake, and have (independently of each other) derived the correct formulas. The details are explained by Darlington (1968, p. 173).

The shortcoming of the corrected formulas is that they apply only to multiple regression techniques of test-construction. As Darlington has observed (1966, p. 322; 1968, p. 175), the multiple-regression technique is not a good test-construction technique for many common situations. We shall therefore ignore these formulas in the present project.

1.52 The Mosier double cross-validation technique

The Mosier (1951) double cross-validation technique is a valuable extension of the traditional cross-validation technique. In the Mosier technique, the total available sample of subjects is split randomly into two subsamples of equal size. A preliminary test is constructed in each subsample. Each of the two preliminary tests is cross-validated in the other subsample. A final test is constructed in the total sample. The mean of the cross-validity figures of the two preliminary tests is used as an estimate of the validity of the final test. The final test is the test published and used in subsequent work.

The double cross-validation technique is more efficient than the usual technique for two reasons. First, the final test is constructed using the entire available sample of subjects; it can thus be expected to have a higher true validity than the usual test based on only a subsample of subjects. Second, all subjects are used, in some way, in estimating the validity of the final test.

The difficulty with the double cross-validation technique, which was recognized by its inventor, is that the technique underestimates the validity of the final test. Nobody knows how serious this underestimation is. The underestimation occurs because the estimate used is the mean of the validities of the two preliminary tests. Since the preliminary tests were based on smaller samples of subjects than the final test was, they can be expected to be less valid than it is. Thus validity figures which apply to the preliminary tests are underestimates for the final test. Similarly, the mean of two such figures (which is the statistic used in the double cross-validation technique) will also tend to be an underestimate.

1.53 The Tukey leave-one-out technique

The Tukey leave-one-out technique (Mosteller & Tukey, 1968) involves constructing a final test plus no fewer than N preliminary tests, where N is the number of subjects in the total sample. Each preliminary test is constructed in the entire sample of subjects, minus one of the N subjects. The one subject left out is different for each preliminary test. For each preliminary test, the investigator then computes the squared error with which the test predicts the criterion score of the one subject not used in the construction of the test. Since there are N preliminary tests, there are altogether N such squared errors. The mean of these N squared errors is used as the estimate of the validity of the final test; it can be translated into a correlation coefficient using well-known elementary formulas.

Although the Tukey procedure involves no appreciable underestimation of the validity of the final test, the procedure is obviously very tedious computationally, even by the standards of modern electronic computers. It involves repeating a complex statistical procedure (construction of a test) N times, where N may well be several hundred. This can easily involve several hundred dollars of computer time, while the standard cross-validation technique of Section 1.32 or the Mosier technique of Section 1.52 usually involve $5 or less of computer time.

2. Description of the new method

The new method is similar to the double cross-validation and leave-one-out methods in that the test to be finally published and marketed is constructed using the data of the total available sample. The methods differ only in the means by which the validity of this final test is estimated.

In the new method, the total available sample of subjects is divided randomly into three subsamples of equal size. A test is constructed in each subsample. These three tests, which we will call X_1, X_2, and X_3, are each constructed by the same method used in constructing the final test. The next step in the method is to estimate the validities and intercorrelations of these three subtests. In order to obtain essentially unbiased estimates of these correlations, each estimate must be computed in a sample not used in constructing the tests being correlated. That is, to estimate the validities of the subtests, each subtest is cross-validated in the two-thirds of the total sample not used in its construction. The three cross-validities thus obtained are denoted by r_{YX_1}, r_{YX_2}, r_{YX_3}, where Y denotes the criterion variable. To estimate the intercorrelations between subtests, each pair of subtests is intercorrelated in the one-third of the total sample not used in the construction of either of the two. That is, X_1 is correlated with X_2 in the subsample used in the construction of X_3. (This subsample, it will be recalled, was not used in the construction of either X_1 or X_2.) This correlation is denoted as $r_{X_1X_2}$. Similarly, $r_{X_1X_3}$ is computed in the one-third of the total sample not used in the construction of either X_1 or X_3, and $r_{X_2X_3}$ is computed in the one-third of the total sample not used in the construction of either X_2 or X_3.

In summary, then, the six correlations r_{YX_1}, r_{YX_2}, r_{YX_3}, $r_{X_1X_2}$, $r_{X_1X_3}$, $r_{X_2X_3}$ are each computed in the subsamples which were not used in the construction of the particular test or tests involved in the correlation being computed.

These six correlations are then entered into the formula

$$(1) \qquad \frac{r_{YX_1} + r_{YX_2} + r_{YX_3}}{\sqrt{3 + 2(r_{X_1X_2} + r_{X_1X_3} + r_{X_2X_3})}}$$

This formula estimates the validity of the final test.

3. Assumptions, derivation, and mathematical properties of the new method.

Though the final argument for Formula 1 must be the data concerning its accuracy in actual problems (reported in a later section), this section gives the mathematical argument which first suggested the formula. It also describes an assumption necessary for deriving the formula, and some mathematical properties of the formula.

3.1 An assumption about the test-construction method

With the exception noted in this section, the derivation in Section 3.2 makes no assumptions about the test-construction technique used or the characteristics of the test constructed. The test score need not be a simple linear sum of item scores; curvilinear and configural item weights are permitted.

The one assumption which is used can be most clearly stated with the help of some additional notation. Let Z_1 be the final test based on the total available sample of people; i.e., the test which is to be published and marketed. Then, as mentioned earlier, X_1, X_2, and X_3 are tests constructed using the same test-construction method used in constructing Z_1, but with each of the three tests based on a different third of the total sample. After adjusting X_1, X_2, and X_3 so that they have the same standard deviation, let Z_2 be the variable formed by averaging, for each person, his scores on X_1, X_2, and X_3. That is,

$$Z_2 = \frac{1}{3}(X_1 + X_2 + X_3).$$

Then the assumption underlying the present technique is that the validity of Z_1 is approximately equal to the validity of Z_2. This is because Formula 1 actually estimates the validity of Z_2, as will be shown in Section 3.2.

3.2 Derivation of Formula 1

The problem, then, is to show that Formula 1 gives an estimate of r_{YZ_2}, the validity of Z_2. By definition,

$$(2) \quad \rho_{YZ_2} = \frac{\mathrm{Cov}(YZ_2)}{\sigma_Y \sigma_{Z_2}} \quad .$$

We can solve for the entries on the right side of (2) as follows. Letting σ denote the common standard deviation of X_1, X_2, and X_3, letting σ_Y denote the standard deviation of Y, and recalling that $Z = \frac{1}{3}(X_1 + X_2 + X_3)$, standard formulas (DuBois, 1965, pp. 215-218) show that

$$(3) \quad \mathrm{Cov}(YZ_2) = \tfrac{1}{3}\mathrm{Cov}\left[Y(X_1+X_2+X_3)\right] = \tfrac{1}{3}\left[\mathrm{Cov}(YX_1)+\mathrm{Cov}(YX_2)+\mathrm{Cov}(YX_3)\right]$$

$$= \tfrac{1}{3}\sigma_Y\sigma(\rho_{YX_1} + \rho_{YX_2} + \rho_{YX_3}),$$

and

$$(4) \quad \sigma^2_{Z_2} = \mathrm{Var}\left[\tfrac{1}{3}(X_1+X_2+X_3)\right] = \tfrac{1}{9}\mathrm{Var}(X_1+X_2+X_3)$$

$$= \tfrac{1}{9}\left[\sigma^2_{X_1}+\sigma^2_{X_2}+\sigma^2_{X_3}+2\cdot\mathrm{Cov}(X_1X_2)+2\cdot\mathrm{Cov}(X_1X_3)+2\cdot\mathrm{Cov}(X_2X_3)\right]$$

$$= \tfrac{1}{9}\sigma^2\left[3+2(\rho_{X_1X_2}+\rho_{X_1X_3}+\rho_{X_2X_3})\right].$$

Substituting (3) and (4) into (2), we see that σ and σ_Y both cancel, and we have

$$(5) \quad \rho_{YZ_2} = \frac{\rho_{YX_1}+\rho_{YX_2}+\rho_{YX_3}}{\sqrt{3+2(\rho_{X_1X_2}+\rho_{X_1X_3}+\rho_{X_2X_3})}}$$

When each population correlation coefficient on the right side of (5) is replaced by the sample correlation coefficient which estimates it, the result is Formula 1, which concludes the derivation. Each estimate entering Formula 1, however, must be a "cross-validation" estimate. The specific sample used in computing each sample correlation coefficient was described in Section 2.

3.3 Mathematical properties of Formula 1

Inspection of Formula 1 shows that the higher the intercorrelations among X_1, X_2, and X_3, the lower is the estimated validity of the final test in relation to the mean of the validities of the three

preliminary tests. If the intercorrelations all equal unity, then Formula 1 reduces to

$$(6)\quad \tfrac{1}{3}(r_{YX_1}+r_{YX_2}+r_{YX_3}),$$

the mean of the validities of the three preliminary tests (which in this case are all equal). The lower the intercorrelations of the three preliminary tests, the higher above (6) is the estimated validity of the final test. This property of the formula is clearly reasonable.. Thus, if the tests constructed in different subsamples intercorrelate very highly, then the final test should be very similar to the three preliminary tests, and it will not be much more valid then they are. On the other hand, if the three preliminary tests have low intercorrelations, then this implies that the smallness of the three subsamples has substantially lowered the validities of the three preliminary tests, so that the final test based on the larger total sample should be substantially more valid than the preliminary tests.

4. An empirical study of the accuracy of the new method

This section describes an empirical study which was done on the accuracy of the validity estimates computed by Formula 1.

4.1 Hypotheses tested

Two specific hypotheses were tested:

1. The expected value of Formula 1 equals approximately the true validity of the final test. The argument suggesting this hypothesis was given in Section 3.2.

2. The standard error of Formula 1 is equal to, or only slightly greater than, the standard error of an ordinary product-moment correlation coefficient ρ. This standard error is known (Anderson, 1958, p. 77) to be approximately equal to $(1-\rho^2)/\sqrt{N}$, where N is the sample size. It seems clear that the standard error of Formula 1 could not be less than this; whether it is greater, or by how much, is one of the questions to be answered by this study.

4.2 Design

A large "population" of 1555 young noncollege employed male high school graduates was used. These data were obtained from Project TALENT. Fifty samples of size 168, and fifty samples of size 84, were drawn randomly from this population. (Each subject was allowed to appear in as many samples as he was drawn for.) In each of these samples, a test was constructed to predict a criterion

variable. The validity of the test was then estimated by applying Formula 1 to the data in the sample in which the test was constructed. The true validty of the test in the total population was then computed. The estimated validity was then compared to the true validity.

The item pool used for constructing the test consisted of 197 items on which the subject rated, on a five-point scale, his interest in various activities like "Swimming" or "Studying". The criterion variable consisted of the sum of eight other items of similar format.

Two different test construction methods were used. One method consisted simply of forming the test from the 50 items with the highest validities. The second method was a modification of the first which attempted to choose items with low intercorrelations in order to increase validity. The method is described in detail by Darlington and Bishop (1966). Both of these methods were used in all 100 samples described above, resulting in 200 tests. The true validity of each of these 200 tests was computed in the total population, and this validity was also estimated from sample data by Formula 1. (Since use of Formula 1 involves constructing three additional tests for each of the 200 tests whose validity is to be estimated, this involved constructing 600 more tests, for a grand total of 800 tests.)

4.21 Deviations from original proposal

The design as described above deviated in the following ways from the design described in the proposal.

1. We originally intended to use salary (for noncollege men) and collegiate grade-point average (for college men) as criterion variables in the study. We were considerably surprised to find that we were unable to construct tests, from the interest items available, which predicted these criteria with much better than zero validity. Since we wished to demonstrate that Formula 1 neither overestimated nor underestimated true validities, and since it is impossible to underestimate a validity if it is zero, we wished to use tests whose true validities were at an intermediate level rather than zero. The easiest way to do this was to use a different criterion variable which was more easily predictable.

2. We originally proposed to study in detail sample sizes of 21 and 42 as well as the sizes of 84 and 168 actually studied. We did some preliminary analyses at these smaller sample sizes, and found the samples simply too small for Formula 1 to have any meaning at all. For example, applying Formula 1 to a sample of size 21 involves constructing tests in three subsamples of size 7 (resulting, as we found, in almost pure random error) and intercorrelating such

tests in other samples of size 7 (producing a correlation which is almost pure random error). Samples of 42, involving subsamples of size 14, were not much better. We therefore did not carry through these analyses in the detail originally proposed.

3. Two other changes were necessitated primarily for economic reasons. This possibility was anticipated and was discussed in the proposal (page 10). We performed only 50 replications at each sample size, rather than the 100 originally suggested. Also we used only one population (noncollege men) rather than the two originally proposed. As it was, we overran our computer budget by over $200, with some of the excess coming from a grant from Cornell and some coming from other budget items in the grant. As anticipated in the proposal, it turned out to be very difficult to predict how much a given analysis will cost. In the final analysis, we constructed 800 tests and analyzed them for $800 in computer time. A cost of $1 per test constructed does not seem excessive, especially since it includes considerable computer cost for "debugging" the programs written.

4. While still intending to use salary as the criterion variable, we selected from the TALENT sample all noncollege full-time-employed high school graduates who had not served in the Army and who had provided salary data. This set of restrictions resulted in a sample of 1555 subjects. When we discovered, to our great surprise, that salary was not predictable from the interest items in our pool, it was too late to prepare a new data set without an additional expense of perhaps several hundred dollars. We therefore went ahead with the sample of 1555 subjects. While somewhat smaller than the set we had originally planned to use, we do not feel the study was affected adversely. The only use of this total sample is to compute test validities, and the only advantage of the larger sample would be to estimate validities more accurately. We feel that a sample of 1555 subjects is still adequate to estimate validities very accurately; the standard error of a correlation coefficient with this sample size is about .02 .

4.3 Results and discussion

The principal results of the study are shown in Tables 1 and 2. Table 1 shows the results pertinent to Hypothesis 1, concerning the expected value of Formula 1. Table 2 shows the results pertinent to Hypothesis 2, concerning the standard error of Formula 1.

The results in Table 1 show that Hypothesis 1 (that Formula 1 gives an unbiased estimate of true validity) is confirmed as strongly as one could imagine it might be. Over a total of 200 tests (50 in each of four conditions), the mean true validities of the tests

exceeded the mean estimated validities of the tests by only .0031. This figure is essentially zero in both a practical and statistical sense; a t test comparing the 200 true validities to the 200 observed validities showed no significant difference between the two means ($t < 1$, $p > .30$). We conclude that Formula 1 can be regarded, for all practical purposes, as an unbiased estimator of a test's true validity.

Table 1

Expected Value of Validity
Estimates Produced by Formula 1

	Mean True Validity*	Mean Estimated Validity	Difference
84 subjects used to construct test and to estimate test's validity			
Test-construction method 1	.4821	.5133	.0312
Test-construction method 2	.5228	.5028	-.0200
168 subjects used to construct test and to estimate test's validity			
Test-construction method 1	.5138	.5089	-.0049
Test-construction method 2	.5643	.5457	-.0186

Mean of four mean differences = -.0031

*Computed in population of 1555 subjects

The results in Table 2 show that the observed standard error of Formula 1 is not grossly larger than the lower limit postulated in Hypothesis 2 (which was $(1-\rho^2)/\sqrt{N}$, the standard error of an ordinary correlation coefficient), but that there is a definite difference between the two. For sample sizes of 84 subjects, the observed standard errors (for test-construction methods 1 and 2 respectively) were 1.59 and 1.63 times the postulated lower limit; both differences

from the lower limit are significant at the 1% level by a chi-square test. For sample sizes of 168 subjects, the observed standard errors were 1.16 and 1.29 times the postulated lower limit; neither difference from the lower limit is significant at the 5% level. From the fact that the observed standard errors for the larger sample sizes were so much closer to their postulated lower limits than they were for the smaller sample sizes (1.16 and 1.29 vs. 1.59 and 1.63), it seems clear that the standard error of Formula 1 rapidly approaches its postulated lower limit as the sample size increases; it seems a reasonable guess that with sample sizes above about 250 or 300, the standard error of Formula 1 would not differ more than about 10% from its theoretical lower limit.

Table 2

Standard Error of Validity
Estimates Produced by Formula 1

	Postulated Lower Limit*	Observed Value	Ratio
84 subjects used to construct test and to estimate test's validity			
Test-construction method 1	.0837	.133	1.59
Test-construction method 2	.0792	.129	1.63
168 subjects used to construct test and to estimate test's validity			
Test-construction method 1	.0567	.0657	1.16
Test-construction method 2	.0525	.0677	1.29

* $(1-\rho^2)/\sqrt{N}$ where ρ is taken from the first column of Table 1.

5. Conclusions

There are three principal extensions of the standard cross-validation method for estimating the validity of an empirically-constructed test: Mosier's double cross-validation method, Tukey's leave-one-out technique, and the present echnique. Tukey's leave-one-out technique is substantially more expensive than the other two, so the choice is often between those. The double cross-validation technique can be presumed to have a standard error of estimate roughly equal to the value postulated above as the lower limit of the standard error of the present technique. The standard error of the present technique is therefore noticeably, but not grossly, above that of the double cross-validation technique. The double cross-validation technique, on the other hand, systematically underestimates the true validity of the test constructed, while the present technique gives an essentially unbiased estimate. The present technique therefore seems to be a reasonable alternative to the double cross-validation procedure, but neither method is clearly superior. More precise rules concerning the uses of the two methods must await further research.

16

6. References

Anderson, T. W. Introduction to multivariate statistical analysis. New York: Wiley, 1958.

Cronbach, L. J. Essentials of psychological testing. New York: Harper, 1960.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Darlington, R. B. Evaluating the use in several schools of assessment procedures developed in a single school. Final report, U.S.O.E. Project No. 3054, 1967.

Darlington, R. B. and Bishop, Carol H. Increasing test validity by considering interitem correlations. Journal of Applied Psychology, 1966, 50, 322-330.

Darlington, R. B. and Stauffer, G. F. Use and evaluation of discrete test information in decision making. Journal of Applied Psychology, 1966, 50, 125-129.

DuBois, P. H. Introduction to psychological statistics. New York: Harper & Row, 1965.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Mosier, C. I. The need and means of cross-validation. Educ. and Psychol. Meas., 1951, 11, 5-28.

Mosteller, F. and Tukey, J. W. Data analysis, including statistics. In G. Lindzey and E. Aronson (Eds.) Handbook of social psychology, Vol. 2, 1968, 80-203.

Wherry, R. J. A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 1931, 2, 440-457.