DOCUMENT RESUME

ED 258 989 TM 850 357

AUTHOR	Livingston, Samuel A.
TITLE	Estimating the Reliability of Classifications Based on Composite Scores.
PUB DATE	19 Nov 84
NOTE	15p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Equated Scores; Essay Tests; *Estimation (Mathematics); Mathematical Models; Scoring; Statistical Analysis; Test Length; Test Reliability; *True Scores; Weighted Scores
IDENTIFIERS	*Composite Scores; *Composite Tests

ABSTRACT

Much previously published material for estimating the reliability of classification has been based on the assumption that a test consists of a known number of equally weighted items. The test score is the number of those items answered correctly. These methods cannot be used with classifications based on weighted composite scores, especially if the composite includes essay scores. This paper presents a modification which will make it possible to apply these methods to composite scores. The proposed method is based on a normal model (with variance stabilizing transformation) for the conditional observed score distribution. The effective test length of the composite is determined from its true-score variance, estimated by Kristof's method or by Gilmer and Feldt's method. (Author/DWH)

November 19, 1984

Estimating the Reliability of Classifications Based on Composite Scores

Samuel A. Livingston
Educational Testing Service

Abstract

Previously published methods for estimating the reliability of classification cannot deal with classifications based on weighted composite scores, particularly if the composite includes essay scores. This paper presents a method based on a normal model (with variance-stabilizing transformation) for the conditional observed-score distribution. The effective test length of the composite is determined from its true-score variance, estimated by Kristof's method or by Gilmer and Feldt's method.

July 30, 1984

Estimating the Reliability of Classifications Based on Composite Scores

Samuel A. Livingston

The problem

Several papers and articles have dealt with the problems of estimating the reliability of classifications based on test scores (e.g., Huynh, 1976; Subkoviak, 1976; Wilcox, 1981; Livingston and Wingersky, 1982). All of these articles are based on the assumption that the test consists of a known number of equally weighted items, scored simply as correct or incorrect, and that the test score is the number of those items answered correctly. This situation is certainly a common one. However, in some testing programs, students are classified on the basis of a composite score -- a weighted sum of scores on two or more tests. The components may not be equally weighted. The component tests may include not only objective tests, but also essay questions. The student's score on each of the essay questions may be a scorer's judgment, expressed on a scale with several possible values. In this case, determining the length of the test for the purpose of estimating reliability is more than a simple matter of counting test items. Can the methods that have been developed for estimating the reliability of classifications be applied when the classification is based on such a composite? The purpose of this paper is to suggest a modification that will make it possible to apply these methods to composite scores.

Notation

Let X_c represent the raw composite score formed from objective component X_0 and essay components X_1, X_2, etc. with weights w_0, w_1, w_2, etc. Let T_c, T_0, T_1, T_2, etc. represent the corresponding true scores.

Then

$$X_c = w_0X_0 + w_1X_1 + w_2X_2 + \dots$$

$$T_c = w_0T_0 + w_1T_1 + w_2T_2 + \dots$$

Let "X_c max" represent the composite score of a student who answers all the objective items correctly and receives the highest possible score on each essay question.

The general method

Many different statistics have been suggested for describing the reliability of classifications based on test scores. These include joint probabilities, conditional probabilities, conditional score distributions, and summary statistics or indices. These statistics are all ways of summarizing the information contained in a joint distribution, and they can be applied to either of two joint distributions: (1) the joint distribution of true scores and observed scores (the "joint T,X distribution"), and (2) the joint distribution of observed scores on alternate forms of the test (the "joint X,X distribution").

If we can estimate the joint T,X distribution (true vs. observed scores), we can use it to estimate the joint X,X distribution (observed scores on alternate forms). The joint T,X distribution gives us both the conditional distribution of X, given T, 'd the marginal distribution of T in the test-taker population. We assume that observed scores on alternate forms are independent, for students with a given true score. This assumption enables us to estimate the joint X,X distribution, conditional on T. We then sum over the marginal distribution of T, to get the joint X,X distribution in the test-taker population.

How, then, do we estimate the joint T,X distribution? We need a model for the true-score distribution (of T) and a model for the conditional observed-score distribution (of X, given T). For our true-score model we can fit a beta distribution (Lord, 1965; Huynh, 1976; Wilcox, 1981) or use the observed-score distribution itself (Subkoviak, 1976), with a transformation to shrink the variance. But what model can we use for the conditional distribution of observed scores? The binomial distribution is a suitable model for a score that is the sum of equally weighted, dichotomously scored items. What kind of model can we use for a composite of the type described above?

A model for the conditional observed-score distribution

One way out of this dilemma is to assume that the conditional observed-score distribution of the composite is similar to that of an all-objective test having the same reliability as the composite. The conditional distribution of observed scores on such a test would be binomial, with parameters n and p, where n is the number of items on the and $p = T/n$. This distribution could be approximated closely by a normal distribution, if the scores were first transformed from X to $X' = 2 \arcsin \sqrt{X/n}$. To apply this model to the composite, first express the composite score X_c as a percentage of its maximum value. Then, apply a variance-stabilizing transformation, to produce the transformed score

$$X'_c = 2 \arcsin\sqrt{X_c/X_c \max} \quad .$$

Assume that the conditional distribution of this transformed score, for students with true composite score T_c, is normal, with mean

$$T'_c = 2 \arcsin\sqrt{T_c/X_c \max}$$

and variance $1/n_c$, where n_c is the effective test length of the composite score X_c. That is, n_c is the length of an objective test having tne same reliability as the composite X_c. To complete the model, we need an estimate of n_c.

Estimating effective test length.

To estimate the effective test length of the composite, we must be willing to make an assumption that may actually be only approximately true. We must assume that the true scores T_0, T_1, T_2, . . . are perfectly intercorrelated. That is, we must assume that if we had perfectly reliable measures of the skills measured by the objective portion and by each essay, these measures would correlate 1.00 with each other.

It follows from this assumption that true scores on the composite will be perfectly correlated with true scores on the objective portion. In general, the standard deviation of true scores on a test is directly proportional to the length of the test. Therefore, we can reasonably define the effective test length of the composite score as the length of the objective portion n_o (which we know), scaled up by the ratio of the true-score standard deviations:

$$n_c = n_0[s(T_c)/s(T_0)]$$

We can estimate $S(T_0)$ by applying a conventional reliability formula (alpha, split-halves, etc.), to produce the estimate

$$S(T_0) = S(X_0)\sqrt{r_0}$$

where r_0 is the reliability coefficient of the objective portion. If the objective scores include a correction for guessing, $S(T_o)$ will be artificially inflated. To correct for this effect, multiply $S(T_o)$ by $k/(k-1)$, where k is the number of answer options per item. In this case, $S(X_o)$ must be computed without changing negative scores to zeroes.

At this point, the missing link in the model is an estimate of $S(T_c)$, the standard deviation of true scores on the composite.

Estimating the true-score standard deviation of the composite.

The problem of estimating $s(T_c)$ is the same as that of estimating the reliability of X_c, since $s(X_c)$ can be observed and $s(T) = s(X)\sqrt{r}$. Kristof (1974) and Gilmer and Feldt (1983) have proposed methods for solving this problem. Kristof's method is simpler but requires that the composite be defined as the sum of exactly three components. If the composite includes more than three components, it is possible to combine components to meet this requirement. If the composite includes only two components, it is necessary to divide one of them, presumably the objective portion, into two parts, creating a composite of three components.

Kristof's formula, applied to a test consisting of an objective component and two essay questions, leads to the estimate

$$s(T_c) = \frac{C_{01}C_{02} + C_{01}C_{12} + C_{02}C_{12}}{\sqrt{C_{01}C_{02}C_{12}}}$$

where C_{01}, C_{02}, and C_{03} are the covariances of the weighted components, i.e.,

$$c_{01} = w_0 w_1 \text{ Cov } (X_0, X_1);$$
$$c_{02} = w_0 w_2 \text{ Cov } (X_0, X_2);$$
$$c_{12} = w_1 w_2 \text{ Cov } (X_1, X_2).$$

When the composite consists of exactly three components, Gilmer and Feldt's estimate is identical to Kristof's. When the composite includes four or more components, Gilmer and Feldt's method is considerably more complex than Kristof's, but also more accurate. Gilmer and Feldt (1983) actually developed two methods. Their method "F2" is simpler to apply than their more complicated method "F1". Only method "F2" will be presented here.

Two modifications of Gilmer and Feldt's formulas are necessary. First, their development does not provide for weighting of the components. Second, their formulas provide a solution for the reliability coefficient, rather than the true-score standard deviation of the composite. With the necessary modifications, their method can be expressed as follows.

Compute the covariance matrix of the weighted components $[c_{ij}]$, with cell entries

$$c_{ij} = \text{Cov}\ (w_iX_i,\ w_jX_j).$$

Let the subscript m indicate the row of this matrix for which the sum of the off-diagonal entries is largest:

$$\sum_{j\neq m} c_{mj} \geq \sum_{j=i} c_{ij} \quad \text{for all } i \neq m. \text{ Define}$$

$$D_i = \frac{(\sum_{j\neq i} c_{ij}) - c_{im}}{(\sum_{j\neq m} c_{mj}) - c_{im}} .$$

Note that $D_m = 1$. The Gilmer-Feldt "F2" estimate of the variance of T_c is

$$s^2(T_c) = \frac{s^2(X_c) - \sum_i c_{ii}}{1 - [\sum_i D^2_i/(\sum_i D_i)^2]}$$

The square root of this quantity provides an estimate of $s(T_c)$.

This estimate of the true-score variance of the composite is the piece that completes the model. It leads to an estimate of the effective test length of the composite. The estimate of effective test length gives us an estimate of the variance in the transformed-normal model for the conditional distribution of observed scores, given true score. We can put this model for the conditional observed-score distribution together with a model for the true-score distribution, to get a model for the joint distribution of true scores and observed scores. With this model and the data from a reasonably large sample of test-takers, we can estimate the joint T,X distribution and summarize it any way we like.

If we want to estimate the joint distribution of observed scores on alternate forms, we can begin by dividing the true-score range into fairly small intervals. (If we have used a Subkoviak-type true-score model, we have already made this partition.) We can then assume conditional independence of the two observed score variables within each true-score interval, and compute the joint X,X distribution for each true score interval. We can then weight each of these joint observed-score distributions by the estimated number of test-takers in the true-score interval and sum over the true-score intervals. The result will be an estimate of the joint distribution of observed scores on alternate forms, which we can summarize any way we like.

References

Gilmer, J. S. and Feldt, L. S. Reliability estimation for a test with parts of unknown length. Psychometrika, 1983, 48, 99-111.

Huynh, H. On the reliability of domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Kristof, W. Estimation of reliability and true score variance from a split of a test into three arbitrary parts. Psychometrika, 1974, 39, 491-499.

Livingston S. A. and Wingersky, M. S. Assessing the reliability of tests used t make pass/fail decisions. Journal of Educational Measurement, 1979, 16, 247-260.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Wilcox, R. R. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 1981, 6, 3-32.

Appendix

Adaptation of Gilmer and Feldt's deviation of coefficient "F2" (Gilmer and Feldt, 1983).

Let $X_c = \sum_i w_i X_i$, with all $w_i > 0$.

Then $T_c = \sum_i w_i\ T_i$.

Assume that each component true score T_i is correlated +1.00 with the composite true score T_c. Then for each T_i there is a constant $a_i > 0$ and a constant b_i such that

$$T_i = a_i T_c + b_i\ .$$

Then

$$T_c = \sum_i w_i (a_i T_c + b_i)$$

$$= T_c \sum_i w_i a_i + \sum_i w_i b_i\ .$$

And

$$Var(T_c) = (\sum_i\ w_i a_i)^2\ Var(T_c)$$

because the w_i, a_i, and b_i are constants. And since all the w_i and a_i are positive,

(1) $$\sum_i w_i a_i = 1.$$

Define

$$c_{ij}\quad Cov(w_i X_i,\ w_j X_j).$$

Then

$$c_{ij} = w_i w_j\ Cov(X_i,\ X_j)$$

$$= w_i w_j \text{ Cov}(T_i, T_j)$$

(because of the independence of errors of measurement)

$$= w_i w_j \text{ Cov}(a_i T_c + b_i, a_j T_c + b_j)$$

(2) $$= w_i w_j a_i a_j \text{ Var}(T_c).$$

Therefore,

$$\text{Var}(X_c) = \sum_i c_{ii} + \sum_{i \neq j}\sum c_{ij}$$

$$= \sum_i c_{ii} + \sum_{i \neq j}\sum w_i w_j a_i a_j \text{ Var}(T_c).$$

Solving for Var(T_c),

(3) $$\text{Var}(T_c) = \frac{\text{Var}(X_c) - \sum_i c_{ii}}{\sum_{i \neq j}\sum w_i w_j \; a_i a_j} .$$

To translate this expression into a usable estimate of Var(T_c), we need to express the denominator in terms of observable quantities. Going back to equation (1) and squaring both sides,

$$(\sum_i w_i a_i)^2 = 1$$

$$\sum_i w_i^2 a_i^2 + \sum_{i \neq j}\sum w_i w_j a_i a_j = 1$$

(4) $$\sum_{i \neq j}\sum w_i w_j a_i a_j = 1 - \sum_i w_i^2 a_i^2 .$$

Substituting (4) into (3),

(5) $$\text{Var}(T_c) = \frac{\text{Var}(X_c) - \sum_i c_{ii}}{1 - \sum_i w_i^2 a_i^2}$$

Consider the sum of the off-diagonal elements in the ith row of the weighted covariance matrix $[c_{ij}]$, minus the kth element:

$$[\sum_{j\neq i} c_{ij}] - c_{ik} \quad .$$

By equation (2), this quantity equals

$$[\sum_{j\neq i} w_i w_j a_i a_j \ \mathrm{Var}(T_c)] - w_i w_k \ a_i a_k \ \mathrm{Var}(T_c)$$

$$= w_i a_i [(\sum_{j\neq i} w_j a_j) - w_k a_k] \ \mathrm{Var}(T_c)$$

$$(6) \qquad = w_i a_i [1 - w_i a_i - w_k a_k] \ \mathrm{Var}(T_c)$$

(because, by equation (1), $\sum_i w_i a_i = 1$).

Let row m be the row of the weighted covariance matrix $[c_{ij}]$ having the largest sum of weighted covariances. Define the index

$$(7) \qquad D_i = \frac{(\sum_{j\neq i} c_{ij}) - c_{im}}{(\sum_{j\neq m} c_{mj}) - c_{im}}$$

for the ith row of the matrix. Notice that D_i is defined entirely in terms of observable quantities and that $D_m = 1$. From equation (6),

$$D_i = \frac{w_i a_i [1 - w_i a_i - w_m a_m] \ \mathrm{Var}(T_c)}{w_m a_m [1 - w_m a_m - w_i a_i] \ \mathrm{Var}(T_c)}$$

$$(8) \qquad = w_i a_i / w_m a_m \quad .$$

Therefore,

$$\sum_i D_i = \sum_i \frac{w_i a_i}{w_m a_m} = \frac{1}{w_m a_m} \sum_i w_i a_i = \frac{1}{w_m a_m} \quad .$$

Therefore,

$$w_m a_m = \frac{1}{\sum_i D_i} \quad . \tag{9}$$

Also, from equation (8),

$$\sum_i D_i^2 = \sum_i \left[\frac{w_i a_i}{w_m a_m}\right]^2 = \frac{1}{(w_m a_m)^2} \sum_i w_i^2 a_i^2 \quad .$$

Therefore,

$$\sum_i w_i a_i = (w_m a_m)^2 \sum_i D_i^2 \quad . \tag{10}$$

Substituting (9) into (10),

$$\sum_i w_i^2 a_i^2 = \frac{1}{(\sum_i D_i)^2} \sum_i D_i^2 \quad .$$

Substituting (10) into (5),

$$\mathrm{Var}(T_c) = \frac{\mathrm{Var}(X_c) - \sum_i c_{ii}}{1 - \left[\frac{\sum_i D_i^2}{(\sum_i D_i)^2}\right]}$$

where $c_{ij} = \mathrm{Cov}\ (w_i X_i, w_j X_j)$ and D_i is as defined in equation (7).