DOCUMENT RESUME

ED 137 368 TM 006 169

AUTHOR Noe, Michael J.; Algina, James
TITLE An Investigation of a Single Administration Estimate of a Criterion- Referenced Reliability Index.
PUB DATE [Apr 77]
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Computer Programs; *Criterion Referenced Tests; Simulation; *Test Reliability; *True Scores

ABSTRACT

Single-administration procedures for estimating the coefficient of agreement, a reliability index for criterion referenced tests, were recently developed by Subkoviak. The procedures require a distributional assumption for errors of measurement and an estimate of each examinee's true score. A computer simulation of tests composed of items that were relatively homogeneous in difficulty for each examinee indicates that the coefficient can be adequately estimated using the binomial error model in conjunction with linear regression estimates of true scores. (Author)

An Investigation of a Single Administration Estimate of a Criterion-Referenced Reliability Index

Michael J. Noe
and
James Algina

Center for Educational Development
University of Illinois, Medical Center

Recently, Subkoviak (1976) developed a single-administration procedure for estimating the reliability of a criterion-referenced test composed of items scored 0/1. The resulting reliability index is termed the coefficient of agreement. The procedure represents an important methodological development for criterion-referenced testing because, in line with suggestions by Hambleton and Novick (1973), the coefficient estimates the proportion of mastery classifications that are consistent on two test administrations, while avoiding the necessity of multiple test administrations. Application of the procedure requires an estimate of each examinee's relative true score (in the sequel simply true score). The true score is defined as the expected value of the proportion correct score. Subkoviak (1976) suggests using linear regression true score estimates, but raises a question about the adequacy of the estimates.

Although it is unlikely that the regression of true score on observed score is precisely linear, the regression function should be monotonically non-decreasing. Therefore, a linear regression function should provide a good approximation to the regression function (Dawes and Corrigan, 1974). In particular, when the true score variance is small, a situation that is common in criterion-referenced testing (Hambleton and Novick, 1973; Popham and Husek, 1969), the approximation of the linear to the true regression function should be quite good. Thus, the use of linear regression estimates may be expected to produce reasonably accurate

Paper presented at the 1977 Annual Meeting of the American Educational Research Association, New York City, April 4-8, 1977.

estimates of the coefficient of agreement.

PURPOSE OF THE INVESTIGATION

In light of the introductory remarks, the purpose of the study was to investigate the accuracy of coefficients of agreement estimated on the basis of three different true score estimates. The first two estimates were obtained for the i^{th} examinee using the linear regression equation

$$[1] \qquad \hat{T}_i = \hat{\beta}\, p_i + (1-\hat{\beta})\, \hat{\mu}_p \qquad (i = 1,2,\ldots,N),$$

with $\hat{\beta}$ set equal to either the sample KR-20 or KR-21 coefficient. The symbol p_i is the observed proportion correct score, $\hat{\mu}_p$ is the sample mean proportion correct score, $\hat{T}_i$ is the estimated true score and $\hat{\beta}$ is an estimate of the slope parameter. The third true score estimate was simply p_i. These three estimates are referred to as the KR-20, KR-21 and proportion correct true score estimates. Once a true score estimate is obtained, an estimate of the coefficient of agreement, P_c, for a given cut-off score, c, can be computed using the formula

$$[2] \qquad P_c = N^{-1} \sum_{i=1}^{N} \{[\text{Prob}(np_i \geq c|T_i)]^2 + [\text{Prob}(np_i < c|T_i)]^2\},$$

with $\hat{T}_i$ estimating T_i and n equal to the number of items. In order to use equation [2] an assumption about the conditional distribution of np_i must be made. Subkoviak (1976) suggests the binominal or compound binominal distribution.

Accuracy of estimation was studied in terms of indices of bias and variability for coefficient of agreement estimators based on the three true score estimates. The accuracy of estimation should be dependent to some extent on the homogeneity of the examinees, the number of items, number of examinees and the cut-off score used to make mastery decisions. The effects of these factors were investigated by a computer simulation of test performance.

DESCRIPTION OF THE SIMULATION

For each of six combinations of number of examinees (N=10,30) and number of items (n=5,10,20), three matrices were constructed with elements p_{ij} (i=1,2,...,N; j=1,2,...,n) representing the true probability of success for the i^{th} examinee on the j^{th} item. These matrices were used in simulating the responses of three groups of N examinees to n items. The true score variance, with true score defined by $T_i = n^{-1} \sum_j p_{ij}$, differed for the three groups. Values of parameters describing the 18 simulated tests are reported in Table 1. The parameters - true score variance, error variance, mean true score and reliability, are defined as

---

Insert Table 1. About Here

$$\sigma^2_T = N^{-1} \left[\sum_i \left(n^{-1}\sum_j p_{ij}\right)^2 - N^{-1} \left(n^{-1}\sum\sum_{ij} p_{ij}\right)^2\right],$$

$$\sigma^2_E = (Nn^2)^{-1} \sum\sum_{ij} p_{ij} (1-p_{ij}),$$

$$\mu_T = (nN)^{-1} \sum\sum_{ij} p_{ij},$$

and

$$\rho^2_{XT} = \sigma^2_T (\sigma^2_T + \sigma^2_E)^{-1}.$$

The computer simulation for each of the 18 tests was accomplished as follows:

1. Generate a N x n matrix of item scores by conducting Nn independent Bernoulli trails. The ij^{th} score takes the value 1 with probability p_{ij} and the value 0 with probability $1-p_{ij}$.
2. From the matrix of item scores compute the three true score estimates for each of the N examinees.
3. Using the three true score estimates in conjunction with the binomial error model[2] compute three coefficients of agreement for each of the n cut-off scores (1,2,3,...,n).[3] These three coefficient of agreement estimators and particular values for each estimator are referred to as the KR-20, KR-21 and proportion correct estimators and estimates respectively.

4. Repeat steps 1-3 for 100 independent replications.

5. Compute deviation statistics (see Tables 2, 3 and 4) over 100 replications for the estimated coefficients.

"True" coefficients of agreement for the n cut-off scores were computed for each of the 18 matrices using the expansion of the compound binomial distribution given in Lord and Novick (1968, p. 525).

CONSIDERATIONS IN CONDUCTING THE SIMULATION

Number of Examinees and Items

Tests lengths of 5, 10 and 20 items were chosen because these values are typical test lengths discussed in the criterion-referenced testing literature (cf, Novick and Lewis, 1974; Hambleton, Hutton and Swaminathan, in press). The numbers of examinees were 10 and 30. These numbers were thought to be representative of typical class sizes and different enough to detect the effects of changing the number of examinees.

Homogeneity of p_{ij}'s

The average within-examinee variance of the p_{ij}'s was small for all matrices, indicating the items are homogeneous in difficulty for each examinee. These p_{ij}'s were chosen to simulate examinee response tendencies to criterion-referenced tests comprised of items that are homogeneous in content. (See Millman (1974) for a discussion of whether criterion-referenced tests must be comprised of items that are homogeneous in content.)

Sampling of Examinees

For each replication the true scores remain the same and therefore estimation of the coefficient for a population of examinees, on the basis of a random sample, is not an issue. Rather, the issue is estimation of a coefficient for a population of administrations of the same test on the basis of data obtained

from a single administration of the test. When a test is used to make decisions on a specific group of examinees, interest should reside in the replicability of the decisions for that group.

Sampling of Items

It is often asserted that criterion-referenced tests should be constructed by following procedures that permit the items comprising a test to be interpreted as a random sample from a well-defined domain of items (cf, Hambleton, Swaminathan, Algina and Coulson, 1974; Millman, 1974). It follows that the coefficient of agreement expected for any two tests constructed by random sampling will be of interest. However, regardless of whether random sampling is actually accomplished, in many instructional contexts only one exam is administered and decisions are based on this administration. Therefore, the coefficient of agreement expected for any two replications of the test (or strictly parallel tests) is also of interest. This simulation focuses on the latter coefficient of agreement and for this reason sampling of items is not an issue.

RESULTS

Statistics summarizing the results of the simulation are reported in Tables 2, 3 and 4. Statistics are not reported for the runs with 10 examinees since the mean deviations for these runs are quite similar to the mean deviations for the runs with 30 examinees. The effects of number of examinees on the variability are discussed in a subsequent subsection. The results based on the KR-20 and KR-21 estimates of true scores typically differ only in the third decimal place and so the latter results are not reported. The existing differences in the mean deviation generally favor the coefficient based on KR-20 true score estimates. The statistics for the cut-off scores not represented in Tables 2, 3 and 4 indicate that the estimates are quite accurate for these cut-off scores.[4]

Insert Tables 2, 3 and 4 About Here

Several notable trends appearing in the data are summarized below.

Effects of Cut-Off Score Changes.

The bias of each coefficient of agreement estimator, as indexed by the absolute mean deviation, tends to be largest for cut-off scores near $n\mu_T$. For these cut-off scores the bias is positive for the proportion correct estimator and negative for the KR-20 estimator. As the deviation between the cut-off score and $n\mu_T$ increases, the following pattern tends to occur for both estimators: The absolute value of the bias decreases until the sign of the bias changes. The absolute value then increases and finally decreases again. Aspects of the pattern occur for all tests, but the pattern occurs most clearly for the 20 item tests.

The variability of the estimator also tends to be larger for cut-off scores near $n\mu_T$ than for cut-off scores at the extremes of the possible observed score distribution. For the cut-off scores near $n\mu_T$ the variability of the KR-20 estimator tends to be larger than the variability of the proportion correct estimator. However, even the variability of the KR-20 estimator for the cut-off scores near $n\mu_T$ is reasonably small. When N=30 the standard deviation reaches a maximum of about .08.

Effects of Reliability

The effects of varying σ^2_T and of varying number of items will be summarized under the single rubric of effects of reliability.

The bias for the proportion correct estimator tends to decrease with increasing ρ^2_{XT}, while the bias for the KR-20 estimator tends to increase with increasing reliability. For almost all cut-off scores on tests with $\rho^2_{XT} < .35$, the bias of the KR-20 estimator is smaller than that of the proportion correct estimator and is quite small in absolute size. For the test with $\rho^2_{XT} = .47$ neither estimator is uniformly less biased. However, on this test the only relatively large biases occur with the KR-20 estimator for cut-off scores equal to seven and eight. For the test with $\rho^2_{XT} = .62$ the proportion correct estimator is less biased for almost

all cut-off scores and the absolute values of the biases are fairly small. In addition, with the exception of cut-off scores 14, 15 and 16, the bias of the KR-20 estimator is also reasonably small.

Effects of Number of Examinees

The bias of the estimators is unaffected by changing the number of examinees. The variability of both estimators increases with the decrease in number of examinees. However, the effect is not very great. When N=10 the maximum observed standard deviation is approximately .10 for the KR-20 estimator.

DISCUSSION

Two of the results deserve further explanation. The first is the change in the sign of the bias as a function of the change in the cut-off score. Consider the idealized situation in which the true score estimates and the true scores have equal means and are linearly dependent. Then for cut-off scores near $n\mu_T$ the coefficient of agreement, calculated using the binomial distribution, will be smaller for the less variable set of numbers. For cut-off scores at the extremes of the possible test scores the coefficient will be larger for the less variable set of numbers. The simulation indicates that

$$[3] \qquad \bar{\hat{\sigma}}^2_T < \sigma^2_T < \bar{\hat{\sigma}}^2_p,$$

where the averages are taken over replications. In [3]

$$\hat{\sigma}^2_T = \hat{\alpha}^2_{20} \quad \hat{\sigma}^2_p,$$

where $\hat{\alpha}_{20}$ is the replication value for KR-20 and $\hat{\sigma}^2_p$ is the estimated proportion correct score variance. Therefore, the KR-20 estimator will tend to underestimate the coefficient of agreement, calculated using $T_i = n^{-1} \sum_j p_{ij}$ in conjunction with the binomial distribution near $n\mu_T$ and overestimate the coefficient for the extreme cut-off scores. In the present study this coefficient is a very close

approximation to the true coefficient, calculated using the compound binomial distribution. Therefore, the KR-20 estimator tends to underestimate the true coefficient near $n\mu_T$ and overestimate the coefficient for the extreme cut-off scores. Moreover, since $\bar{\hat{\sigma}}^2_p > \sigma^2_T$ the opposite relationship holds for the proportion correct estimator.

The second result requiring explanation is the relationship between the bias of the two estimators and ρ^2_{XT}. An explanation relevant to cut-off scores near $n\mu_T$ is offered below. A similar explanation can be extended to other cut-off scores, but in view of space limitations the extension is left to the reader.

For the KR-20 estimator the keys to the explanation are that (1) the smallest possible estimated coefficient of agreement is .50, a value that can occur only when the estimated KR-20 = 0.00, and (2) the KR-20 estimator tends to underestimate the true coefficient for cut-off scores near $n\mu_T$. As ρ^2_{XT} approaches zero the true coefficient approaches .50 for cut-off scores near $n\mu_T$, and therefore the underestimation cannot possibly be great. On the other hand, when ρ^2_{XT} is large the true coefficient can be substantially larger than .50, and the underestimation can be substantial. In Table 2 the mean deviations for cut-off scores 14 and 15 on examinations one and three illustrate these relationships. (The reader should note that the reported statistics or parameters for a particular cut-off score on examinations one and two or two and three are not comparable, since μ_T for exam two differs from μ_T for the other two examinations.) For the proportion correct estimator the keys are that (1) this estimator tends to overestimate the true coefficient of agreement, and (2) the true score distribution is estimated by the observed score distribution. The degree of overestimation will depend in part on the proportion of the estimated true score variance, here the observed proportion correct score variance, that is error variance. When ρ^2_{XT} is low, this proportion is high and overestimation tends to be great. On the other hand, when ρ^2_{XT} is large the degree of error score variance is smaller and therefore the overestimation is smaller.

The relationship between ρ^2_{XT} and the two estimators suggests that when ρ^2_{XT} is large, say greater than .50, the proportion correct estimator might be used. However, it should be noted that KR-20 is quite variable over replications and may be a poor guide to the choice of estimator. A better strategy may be to average the proportion correct and KR-20 coefficients of agreement when KR-20 is large.

CONCLUSION

The results indicate that with few exceptions accurate estimation of the coefficient of agreement can be obtained using the KR-20 estimate of true score in conjunction with the binomial error model, at least for tests comprised of items that are homogeneous in difficulty for each examinee. The coefficients estimated on this basis were substantially biased only for cut-off scores near $n\mu_T$ for tests with $\rho^2_{XT} \geq .47$. Moreover, the variability of the estimator was reasonably small in all cases.

Footnotes

1. This research was funded by a faculty grant from the Center for Educational Development, University of Illinois, Medical Center.
2. As Subkoviak (1976) indicates, the compound binomial is probably a more realistic model for errors of measurement. Initial runs with both models indicated that there was very little difference in the accuracy of estimated coefficients based on the two models, and therefore the cost of duplicating computations was avoided. The similarity is due to the fact the p_{ij}'s are relatively homogeneous for each examinee.
3. From the point of strong true score theory, if the appropriate model for error of measurement is binomial, then the regression parameter should be KR-21. However, when it is desired to estimate the proportion of mastery classifications that will be consistent for repeated administrations of the same or strictly parallel tests, KR-20 provides the better lower bound estimate of the reliability of the test (Lord and Novick, 1968) and probably should be used even if the binomial distribution is employed for the sake of computational convenience.
4. A copy of tables reporting the entire set of results is available from the authors.

References

Dawes, R.M. and Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Hambleton, R.K. and Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R.K., Swaminathan, H., Algina, J., and Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Laboratory of Psychometric and Evaluative Research Report No. 12. Amherst, MA: School of Education, University of Massachusetts, 1974.

Hambleton, R.K., Hutton, L., and Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. Journal of Experimental Education (in press).

Lord, F.M. and Novick, M.R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Co., 1968.

Millman, J. Criterion-referenced measurement. In W.J. Popham (Ed.), Evaluation in Education: Current Applications. Berkeley, CA: McCutchan Publishing Co., 1974.

Novick, M.R. and Lewis, C. Prescribing test lengths for criterion-referenced measurement. In C.W. Harris, M.C. [illegible], and W.J. Popham (Eds.), Problems in Criterion-Referenced Measurement. Los Angeles: Center for the study of Evaluation, University of California, Los Angeles, 1974.

Popham, W.J. and Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

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

Table 1

Parameters Describing 18 Simulated Tests

Matrix Dimensions (examinees x items)	Examination	Parameters			
		σ^2_T	σ^2_E	μ_T	ρ^2_{XT}
(30x20)	1	.0018	.0096	.71	.16
	2	.0039	.0084	.76	.32
	3	.0148	.0092	.70	.62
(30x10)	4	.0029	.0192	.70	.13
	5	.0055	.0169	.75	.24
	6	.0165	.0185	.69	.47
(30x5)	7	.0047	.0389	.70	.11
	8	.0051	.0336	.76	.13
	9	.0170	.0367	.70	.32
(10x20)	10	.0017	.0096	.71	.15
	11	.0036	.0082	.76	.31
	12	.0158	.0092	.70	.63
(10x10)	13	.0022	.0195	.70	.10
	14	.0061	.0165	.76	.27
	15	.0166	.0183	.69	.47
(10x5)	16	.0035	.0381	.71	.08
	17	.0070	.0314	.78	.18
	18	.0144	.0375	.69	.28

Table 2

Indices of Bias and Variability for Two Coefficient of Agreement Estimators: n=20, N=30

Examination	Parameter and Statistics	Cut-Off Scores 9	10	11	12	13	14	15	16	17	18	19
1	True Coefficient	.989	.965	.912	.818	.694	.586	.556	.635	.780	.909	.977
ρ^2_{XT} = .16	Mean Deviation	-.042 .001	-.059 .002	-.060 .002	-.027 .002	.041 -.015	.112 -.031	.134 -.032	.084 -.015	-.004 .001	-.058 .005	-.056 .002
	Standard Deviation	.016 .006	.021 .014	.024 .028	.025 .044	.025 .048	.025 .031	.026 .023	.027 .044	.026 .042	.022 .024	.018 .008
2	True Coefficient	.995	.983	.956	.902	.817	.715	.632	.612	.674	.794	.911
ρ^2_{XT} = .32	Mean Deviation	-.023 .002	-.036 .006	-.044 .013	-.038 .019	-.008 .014	.040 -.014	.082 -.056	.090 -.072	.052 -.042	-.008 .004	-.045 .002
	Standard Deviation	.012 .014	.016 .008	.020 .016	.024 .030	.027 .047	.028 .055	.027 .043	.025 .031	.024 .048	.026 .046	.025 .025
3	True Coefficient	.920	.873	.827	.791	.766	.751	.745	.753	.781	.839	.918
ρ^2_{XT} = .62	Mean Deviation	-.015 .027	.002 .035	.017 .026	.025 -.004	.027 -.044	.026 -.076	.025 -.085	.023 -.060	.013 .016	-.002 .025	-.025 .027
	Standard Deviation	.020 .025	.021 .030	.022 .030	.022 .030	.025 .038	.030 .047	.032 .047	.029 .036	.023 .031	.023 .033	.023 .023

Note: For the rows corresponding to each statistic, the first line is for the proportion correct estimator and the second is for the KR-20 estimator.

Table 3

Indices of Bias and Variability for Two Coefficient of Agreement Estimators: n=10, N=30

Examination	Parameter and Statistics	Cut-Off Scores 4	5	6	7	8	9	10
4	True Coefficient	.977	.909	.760	.589	.561	.740	.936
ρ^2_{XT} = .13	Mean Deviation	-.055 -.001	-.056 -.002	.013 -.009	.118 -.022	.131 -.022	.012 -.007	-.059 .000
	Standard Deviation	.020 .012	.028 .033	.029 .057	.027 .045	.025 .029	.027 .054	.025 .024
5	True Coefficient	.986	.945	.844	.694	.590	.657	.866
ρ^2_{XT} = .24	Mean Deviation	-.037 .004	-.050 .011	-.021 .012	.055 -.017	.120 -.052	.074 -.032	-.019 .010
	Standard Deviation	.031 .019	.044 .042	.050 .070	.049 .083	.058 .066	.063 .074	.053 .060
6	True Coefficient	.924	.838	.751	.700	.690	.745	.891
ρ^2_{XT} = .47	Mean Deviation	-.028 .024	.001 .028	.031 -.027	.044 -.094	.049 -.086	.038 .007	-.007 .034
	Standard Deviation	.021 .026	.024 .039	.030 .050	.031 .057	.032 .050	.032 .046	.027 .031

Note: For the rows corresponding to each statistic, the first line is for the proportion correct estimator and the second is for the KR-20 estimator.

Table 4

Indices of Bias and Variability for Two Coefficient of Agreement Estimators: n=5, N=30

Examination	Parameter and Statistics	Cut-Off Scores 2	3	4	5
7	True Coefficient	.936	.737	.542	.724
	Mean Deviation	-.056	.040	.165	.073
ρ^2_{XT} = .11		-.002	-.005	-.016	-.010
	Standard Deviation	.024	.031	.033	.033
		.025	.056	.032	.052
8	True Coefficient	.969	.828	.589	.643
	Mean Deviation	-.056	-.014	.144	.156
ρ^2_{XT} = .13		-.004	-.004	-.013	-.015
	Standard Deviation	.024	.032	.036	.035
		.019	.054	.055	.052
9	True Coefficient	.906	.752	.622	.715
	Mean Deviation	-.035	.029	.109	.100
ρ^2_{XT} = .32		.017	-.022	-.075	.001
	Standard Deviation	.023	.027	.032	.036
		.029	.051	.040	.046

Note: For the rows corresponding to each statistic, the first line is for the proportion correct estimator and the second is for the KR-20 estimator.