DOCUMENT RESUME

ED 075 513 TM 002 616

AUTHOR	Mandeville, Garrett K.
TITLE	Confidence Interval Estimation of KR sub 20--Some Monte Carlo Results.
PUB DATE	Feb 73
NOTE	18p.; Paper presented at the Annual Meeting of National Council on Measurement in Education and American Educational Research Association inew Orleans, Louisiana, February 26-28, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Confidence Testing; *Item Sampling; *Literature Reviews; Sampling; *Statistical Analysis; Test Reliability
IDENTIFIERS	Monte Carlo Results

ABSTRACT

An investigation is conducted which presents extensive Monte Carlo results which indicate the conditions under which a procedure using the F distribution can be used to study the robustness of the confidence interval procedures for small samples. A review of the literature is presented. Procedure uses a binary data matrix. Results indicate that the procedure is an extremely practical one. (CK)

26.16

CONFIDENCE INTERVAL ESTIMATION OF KR_{20}-- SOME MONTE CARLO RESULTS

Garrett K. Mandeville
University of South Carolina

Paper presented at the annual meeting of the
National Council on Measurement in Education
New Orleans, February 26-28, 1973

CONFIDENCE INTERVAL ESTIMATION OF KR_{20}--

SOME MONTE CARLO RESULTS

Garrett K. Mandeville
University of South Carolina

Introduction

Educational researchers are generally aware of the fact that, unless the measurements used to draw inferences in the study are of sufficient reliability, these inferences may well be meaningless manifestations of random variation. Thus, for standardized tests normative information as presented in the test manual will be cited whereas for instruments which the researcher has constructed, internal consistency reliability coefficients such as Cronbach's coefficient alpha ($\underline{r}_\alpha$) (Cronbach, 1951) or its form when items are scored dichotomously, the Kuder-Richardson 20 ($\underline{r}_{20}$) (Kuder and Richardson, 1937) are frequently presented.

Only rarely, however, do researchers concern themselves with the fact that an instrument does not have a single reliability but that this index is also a function of the population tested. We find, for example, that some standardized tests which are quite reliable when used for measuring middle-class children are virtually useless with Head Start populations.

It occurs to this writer that similar phenomena may be operating in situations where deviations from standard teaching methods or other variations in treatments used or populations sampled, may cause normative information supplied in a test manual to be wholly irrelevant. In educational experiments or quasi-experiments, then, it is the feeling of this writer that adequacy of test reliability should not be taken for granted but should be constantly checked and that this should be done separately for samples which differ on manipulated independent variables.

Since, in many research studies, moderately small samples are being gathered, point estimates of a reliability index do not provide enough information to the researcher who is concerned about whether the instrument 1) is reliable enough for his purpose, (if he has just constructed it), 2) is operating as reliably as reported in the manual (if it is a standardized test) or 3) exhibits consistency of reliability for different groups being tested. It is the feeling of this writer that the simple device of presentation of a confidence interval estimates of the reliability for each experimental group of subjects used in the study would be very useful data to include in the reporting of research results. If the argument developed above is logical, the next question to be addressed is: "What procedures to recommend for confidence interval estimation of the reliability when samples are small"?

Restricting the discussion to inferences about ρ_{20}, the population value of the Kuder-Richardson 20 reliability coefficient, little information will be found on this topic in the literature. The commonly accepted procedure, which utilizes the F distribution, lacks empirical or analytic support when the samples are small. The only other procedures for making inferences about ρ_{20} which this writer has found, were given by Payne and Anderson (1968). These investigators empirically derived an extensive set of tables for testing that ρ_{20} is zero but, unfortunately, they cannot be used for interval estimation. Thus, a study of the robustness of the confidence interval procedures for small samples appears to be the most reasonable first step in any attempt to solve this problem. It is the goal of this investigation to present fairly extensive Monte Carlo results which will indicate the conditions under which this procedure can be used.

The Literature

Feldt (1965, 1969) has presented derivations based on the two-factor random model of analysis of variance (ANOVA) which provide tests of hypotheses and confidence intervals in the one sample case and tests of hypotheses in two sample problems involving $\underline{r}_\alpha$. In the first paper, Feldt clearly points out the problems which arise in using this model to describe dichotomous test item data. Assumptions which are obviously violated are those of normality, homoscedasticity of errors, and independence of the subject effects and errors.

Another problem area is the fact that in common testing procedure a fixed test is used. Thus, the two-factor model is not strictly appropriate, the sampling being Type 1 (Lord, 1955) as Feldt has also pointed out. The application of these procedures to dichotomously scored, fixed test item data might then be considered suspect but, by and large, the impression obtained from the literature is that, because of the well-known robustness of ANOVA procedures, useful results can be obtained.

Although Feldt did present some empirical results which were in general agreement with the theoretical predictions, they were very limited. Using data from a study by Baker (1962), Feldt obtained the distribution of 200 $\underline{r}_{20}$ values for samples of 15, 30 and 60 subjects. The empirical percentiles of the distribution of $\underline{r}_{20}$ compared favorably with those derived from the F distribution.

Until a recent article by Nitko and Feldt (1969), this writer could find no results which considered the effect of item difficulties on the distribution of $\underline{r}_{20}$. Nitko and Feldt, however, showed that the sampling distributions of $\underline{r}_{20}$ are similar for two different distributions of item difficulty and that this was true for five tests with ρ_{20}'s ranging from .55 to .86. For the thirteen item tests simulated, the item difficulty distributions were concentrated around .5 or spread evenly over the range .2 to .8. Although exhibiting the similarity of the two distributions, the results given in the Nitko and Feldt paper do not allow a straightforward comparison of the empirical results with those expected from the F distribution. When this is done, it can be seen that the lower empirical percentiles are slightly larger than those predicted from the F distribution for ρ_{20} larger than .5. This means that there is a deficiency of small values of r_{20}. In Table 1, which follows, are some comparisons of the empirical percentiles of r_{20} presented by Nitko and Feldt and those expected on the basis of normal theory.

- -

Table 1 About Here

- -

Lack of substantial evidence that the F distribution provides a useful model for estimation of ρ_{20} with moderate sized samples caused the present writer to undertake the research presented in this paper. In light of the distributional problems confronted in attempting an analytic solution in the small sample case, a Monte Carlo investigation was undertaken.

Description of the Tests Simulated

One of the ways that tests typically vary, and therefore a useful parameter to consider in a simulation study, is the distribution of item difficulty. In the study presented here, the following three distributions were considered: homogeneous with difficulty parameters from .3 to .7; heterogeneous with difficulties from .1 to .9; and homogeneous with difficulties ranging from .1 to .5. In the discussion to follow, these tests will be abbreviated as HOM, HET, and HARD, respectively. The actual difficulty indices used for ten item tests are given in Table 2. Twenty and thirty item tests were simulated by using two or three items at each difficulty level.

- -

Table 2 About Here

- -

In this study, ρ_{20} was not taken as a parameter. Instead, an approach which assumed that the binary response vector was obtained by partitioning a multidimensional space and applying this partition to a multivariate normal continuous vector of "latent variables" was used. This data generation model is consistent with the popular normal ogive scaling model described elsewhere (e.g.,Lord and Novick, 1968, p. 365-373). Once the success proportions had been designated, the other quantity needed in this data generation scheme was the matrix of intercorrelations of the latent variables associated with the dichotomous item responses. Three matrices were used in the main body of the study and all three were patterned i.e., all pairs of latent variables had the same intercorrelation. These constant correlations were taken to be .1, .3, and .6. The combination of the three correlation structures and three difficulty distributions led to nine test structures. These nine test structures were increased to 27 tests actually simulated by considering tests of 10, 20, and 30 items each and the range of ρ_{20} for these 27 tests was .36 to .95. Since the main concern was the distribution of $\underline{r}_{20}$ for small samples, data for 30 subjects were simulated throughout the study. In order to simulate some actual tests, additional runs were made with four tests described by Ross (1966) and which ranged from 12 to 18 items in length. These tests, referred to as W, X, Y, and Z in the Ross paper, were simulated by using the item difficulties which were given and obtaining the item intercorrelation matrix from the vector of factor loadings of each item on the common factor. The ρ_{20} for each test was larger than .90 and the item difficulties were typically in the .3 to .7 range. The utilization of item parameters which characterized actual tests was felt to be important because of the difficulty in generalizing from the constant correlations used in the rest of the study.

Procedures

Let the assumption be made that a binary data matrix is available representing the responses of $\underline{N}$ subjects to $\underline{k}$ items. $\underline{MS}_S$ and $\underline{MS}_{IxS}$ will refer to the mean squares in the ANOVA corresponding to the subject and item by subject interactions. The quantity $\underline{F}_{ob} = MS_S/MS_{IxS} = (1-\underline{r}_{20})^{-1}$ is then readily computed. The population analogue of $\underline{F}_{ob}$ will be referred to as $\underline{F}_{pop}$ in accord with earlier notation of Feldt (1965) and is related to ρ_{20} by $\underline{F}_{pop}=(1-\rho_{20})^{-1}$. The statistic used in the investigation was $\underline{V}=\underline{F}_{ob}/\underline{F}_{pop}$. The computation of $\underline{F}_{pop}$ was carried out by using the correlations in the latent variable correlation matrix and the item success proportions in a series expansion (Kendall & Stuart, 1961) relating the correlation in the bivariate normal to its phi coefficient.

If the two-factor random model is appropriate, Feldt has shown that $\underline{V}$ should be distributed according to the $\underline{F}$ distribution with $\underline{N}$-1 and ($\underline{N}$-1)($\underline{k}$-1) degrees of freedom. Thus, values of this statistic were cast into a frequency distribution with boundaries which were the deciles of the appropriate $\underline{F}$ distribution. In addition, 90% and 95% open-ended (lower) and closed confidence limits were obtained according to standard procedures derived by Feldt.

For clarification, consider the following probability statements which serve as the basis for the confidence intervals:

(1) $\underline{Pr}(\underline{C}_1 < \rho_{20}) = 1-\alpha$ where $\underline{C}_1 = 1-(1-\underline{r}_{20})\underline{F}_{1-\alpha}$

(2) $\underline{Pr}(\underline{C}_{2L} < \rho_{20} < \underline{C}_{2H}) = 1-\alpha$ where $\underline{C}_{2L} = 1-(1-\underline{r}_{20})\underline{F}_{1-\alpha/2}$

and $\underline{C}_{2H} = 1-(1-\underline{r}_{20})\underline{F}_{\alpha/2}$

Note that (1) and (2) refer, respectively, to open-ended and closed confidence intervals which are often of interest for ρ_{20}. For each new sample generated the three boundary points $\underline{C}_1$, $\underline{C}_{2L}$, and $\underline{C}_{2H}$ were computed for each of α=.10 and .05. Counters were advanced if any of the inequalities presented in probability statements (1) or (2) were violated. These frequency counts were later converted to sample proportions for comparison with the theoretical probabilities. In tables to follow, these three empirical proportions are denoted as $\underline{E}_1$, $\underline{E}_{2L}$, and $\underline{E}_{2H}$* respectively, and the sum of the last two is simply $\underline{E}_2$. One thousand data sets were generated for the ten item tests, 500 for the 20 and 30 item tests. For the four tests from the Ross study, which ranged from 12 to 18 items in length, 1000 data sets were generated.

The population ρ_{20}'s and the average $\underline{r}_{20}$ for the 500 or 1000 values generated are presented in Table 3 along with sample estimates of the skewness and kurtosis of the test score distributions. Summary statistics for the overall fit of the empirical and theoretical distributions are also given in Table 3 in terms of χ^2 goodness of fit statistics. These were computed using the ten categories based on the deciles of the appropriate F distribution.

- -

Table 3 about here

- -

It is the writer's opinion that although the results for short tests where the latent item intercorrelation is low (.10) are not of much practical interest, that for the majority of the tests simulated, the test parameters are similar to those obtained in practical testing situations in education. For example, the symmetric score distributions (HET and HOM) exhibit varying degrees of platykurtosis as is commonly found in achievement and aptitude test score distributions. Exceptions may be the HET test with ρ=.60 which is nearly rectangular, actually slightly U-shaped. Similarly, the skewed distribution for the HARD test with ρ=.6 is a rather severe J-shaped distribution which would be uncommon in most educational

* See footnote to Table 4 for interpretation of these proportions.

settings.

Again referring to Table 3, we observe that the well-known negative bias of $\underline{r}_{20}$ as an estimator of ρ_{20} (e.g., see Lord and Novick, 1968) is not very serious. The average value of $\underline{r}_{20}$ for the samples generated is typically slightly smaller then ρ_{20} when that parameter is small, but the bias becomes trivial when ρ_{20} is larger than about .7.

As regards the χ^2 statistics reported in Table 3, the writer does not view their significance as particularly important, but they are presented to indicate in general how well the distribution of $\underline{V}$ approximates the $\underline{F}$ distribution. For nine of the 27 tests simulated, the goodness of fit statistic was significant at the 5% level, indicating gross lack of fit of the empirical to the theoretical distribution. As the reader surely realizes, what is more important is the fit in the tails of the distributions since this governs the adequacy of the inferential procedures. Comments concerning the χ^2 results, then will be included with the discussion on the accuracy of the confidence interval estimation, the results of which, for the main body of tests simulated, follow in Table 4.

Table 4 about here

RESULTS

In evaluating the results of this investigation, it is useful to consider the sampling variation which can be expected when the binomial distribution is the appropriate model. Presented in the table below are the standard errors for a sample proportion from populations with π=.10 and .05 and based on samples of size 500 and 1000.

Brief table of standard errors of proportions

		Proportion .10	Proportion .05
sample size	500	.013 (.013)	.0098 (.010)
	1000	.0095(.010)	.0069 (.007)

The values in parentheses above are rounded versions which were used to determine intervals within which a resulting proportion might be expected to fall about 68% of the time if the theoretical percentiles were correct. When the empirical proportions F_1 and F_2 in Table 4 are compared with these intervals, it is found that all of the MET tests for ρ=.1 are within the limits imposed. The more platykurtic MET tests with ρ=.3 and .6 have a rather large number of entries, actually

19 of the 24, which are outside of these limits. The interesting fact is that all of these 19 empirical proportions are below the nominal values. Thus for these tests the nominal confidence coefficient tends to underestimate truth, e.g., more than 95% of the "95% confidence intervals" cover the true parameter. When percent relative error, defined as the absolute error divided by the nominal α, is considered it is found to vary from 14% for .90 confidence coefficient, ρ=.3, open interval (E_1 = .086) to 62% for .95 coefficient, ρ=.6 (E_2 = .019). Naturally, percent relative errors are larger for 95% than 90% nominal intervals and, excluding the poorly behaved results for the 10 item HET test with ρ=.6, generally are below 40%. It is worthy of note that for each of the five HET tests with significant χ^2 values, E_{2H} exceeds E_{2L} indicating a shortage of low values of $\underline{V}$ (and $\underline{r}_{20}$). (There is one exception to this trend for ρ=.6, k=30, .95 confidence coefficient). This fact is in keeping with the remarks made earlier in reference to Feldt's findings which suggested that the lower percentiles of the empirical distributions were slightly larger than those for the comparison F distribution. In the four significant χ^2 values for ρ=.3 and .6, the largest contribution to the χ^2 is the contribution from the lowest category. There is no perceptable relation between test length and the adequacy of the estimation procedures.

Moving to discuss the HOM tests, we find that some empirical proportions exceed the one sigma limits at all levels of item intercorrelation and deviate in both directions from the nominal values. Of the nine values of E_1 which were "significant", seven exceeded the nominal value indicating true confidence less than nominal for these open-ended intervals. With the exception of the .95 confidence interval for ρ=.1 and 30 items (E_1 = .070), the other relative errors were 24% or less for these intervals, indicating, for example, that generally no fewer than 94% of the 95% intervals generated covered the true parameter. Thus, although at variance with the conservatism associated with the estimation procedure for the HET tests, the results for open interval estimation for tests satisfying the HOM model still appear to have practical implications.

For closed intervals, 8 of the 18 values of E_2 were outside of the one sigma limits, and, contrary to the results for E_1, seven of the eight yielded "significantly too many" intervals which covered the true parameter. The relative errors showed a definite increase as the item intercorrelation increased and were rather large (34% and 44%) for the test with ρ=.6. It will be recalled that the score distribution for this test is virtually rectangular, however. Examination of the E_{2H} and E_{2L} entries indicates that where any differences between these two figures exist, E_{2H}, which represents the proportion of times that the interval totally exceeds the true parameter, is usually larger than E_{2L}. This is reasonable from the results for E_1 and indicates that the empirical distributions tend to have too much density in the upper tail and too little in the lower tail.

For the HARD tests, the results for open intervals are similar to those for the HOM tests in that all 13 "significant" values of E_1 were larger than the nominal values. However, whereas for the HOM tests the relative errors were usually smaller than 20%, for the HARD test they range up to 68% for nominal .95 coefficient, 30 item test with ρ=.3 (E_1=.084). The extreme J-shaped score distribution for ρ=.6 provides too few "proper intervals" for each of the six combinations of confidence coefficient and number of items and the relative errors appear to increase with the length of the test. Closed intervals for the HARD test are somewhat better behaved for the more practical situations of ρ=.1 and .3. The largest relative error among the six E_2 values which were more than one sigma from the nominal value was 34% which occured for the same simulation as the 68% for the open interval. As a matter of fact, the .084 proportion of overestimates of ρ_{20} combined with precisely the correct number of underestimates (.050) to yield E_2 = .134. For ρ=.6, the relative errors are rather large (22% to 58%) and reflect too few intervals covering the true value. The primary reason is excessive values of E_{2H}. On the other hand, the lower tail of the $\underline{V}$ distribution appears to fit the $\underline{F}$ distribution quite well. The significant χ^2 values of 17.6 for the 30 item HARD tests with ρ=.3 and ρ=.6 are primarily due to the excess of observations in the top category; in each case , the contribution from those categories provided the largest contribution to χ^2.

The combined results of the four tests from the Ross paper follow in Table 5.

Table 5 about here

We observe that tests W and X are quite homogeneous with respect to difficulty, the σ_π values being much smaller than for the HOM and HARD tests simulated. Test Z, on the other hand, has an item difficulty spread similar to these two test models. The average latent item intercorrelation for all four tests is larger than .6 and the largest value of .76 characterized test X. The strong inter-item associations cause all ρ_{20}'s to be above .90. The test distributions for these four are interesting and will be related to the simulated tests already discussed. The easiest one to compare is test Y which is similar in form to the ρ=.60, HOM test except that it is slightly more platykurtic (in this case more U-shaped). When a comparison is made against the results for the 10 item test in this cell, they are found to be very similar. Open intervals do not cover the parameter as often as the nominal coefficient advertised while a shortage of entries in the lower tail caused closed interval construction procedures to be on the conservative side.

The remaining three tests are moderately negatively skewed. In terms of skewness and kurtosis, test W and Z appear similar, but the score distribution for test Z is somewhat more rectangular. Neither of these two distributions has an interior mode.

The results for test Z follow the same general lines of those for the 20 item HOM test with $\rho \bar{=} .6$, i.e., E_1 values are a little too large and E_{2L} too small. It would seem as though the corresponding 10 item test would be useful for comparing to test W, but it soon becomes evident that test W along with test X yield the strongest negative findings in the study. Relative errors of as much as 100% (actually slightly larger) exist for these two tests. Although quantitatively much more deviant, the results follow the general trend of the highly correlated HOM and HARD tests, namely that there are too many values in the upper tail of the V distribution and too few in the lower. The test X score distribution is U-shaped and very extreme.

It appears as though the selection of real tests to simulate may not have been particularly well chosen. The rationale for selecting these was one of easy availability· the information necessary for the generation scheme utilized was readily available. Upon looking at the input values for the four Ross tests, the only parameters which varied between tests W and Z was that of difficulty distribution. Because of this the writer decided to make simulations for tests with all items of the same difficulty. These runs were made simulating the ten item tests with $\rho = .6$ and yielded the results given in Table 6 below.

TABLE 6
Results for π =Constant

π	α	E_1	E_2	E_{2H}	E_{2L}	ρ_{20}	$\bar{r}_{20}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
.5	.10	125	103	066	037	.87	.87	.00	-1.35
	.05	066	053	035	018				
.3	.10	146	155	096	059	.87	.86	.81	- .55
	.05	096	081	055	026				
.1	.10	213	382	160	222	.83	.78	2.38	5.67
	.05	160	311	127	184				

For $\pi = .5$ the test score distribution is U-shaped similar to test Y from the Ross paper and the HOM test for $\rho = .3$. Relative errors for open intervals are around 30%. The second test simulated, with $\pi=.3$, generated a score distribution similar to that of test W and the confidence interval results for these two tests are very similar. Open intervals have relative errors approaching 100% and the over populated upper tail caused the closed intervals to be in error between 50% and 60% more often than the nominal coefficient would suggest. The situation becomes much worse for the very difficult test with $\pi=.1$. The test score distribution is extreme, however, with 64% of the "total scores" generated being zero.

DISCUSSION

The writer set himself to the task of determining the extent to which interval estimation of ρ_{20} using standard procedures based on the F distribution could be relied on for moderate sized samples (N=30). Results for tests with items spread evenly over a wide range of difficulty and which, therefore, resulted in a symmetric test score distribution were in good agreement with "nominal" results. For tests in which items were strongly associated, test score distributions were platykurtic and the nominal confidence coefficient typically underestimated the true proportion of correct statements. Most statisticians find this conservative approach at least tolerable. When the items were spread over a narrower range of difficulty, but were still centered at .5, there was a tendency for too few open intervals to be "correct". The relative errors, however, were small, generally less than 24%. For closed intervals the conservative nature of the HET tests reappeared. Results for test Y from the Ross paper and the tests simulated with $\pi=.5$, both of which had symmetric score distributions were in agreement with these results. Therefore, when the test score distribution was symmetric, the most serious results were in the direction of conservative procedures. The fact that fewer than the nominal % of the open intervals covered the true parameter for the highly associated HOM tests does not seem too serious in that the % error was generally small.

In the situations simulated where the score distribution was skewed, it was virtually always true that too few open intervals covered the true parameter and % errors ranged up to and sometimes exceeded 100%. The most severely skewed score distributions, with no interior mode, occurred for the HARD tests with $\rho=.6$, three of the four Ross tests and the tests with constant difficulty parameters of .3 and .1. A somewhat conservative rule which could be used for open intervals in these cases would be to use the 97.5th percentile of the F distribution to construct open 95% intervals. The only situation where such an adjustment procedure would not be either conservative or within reason was the $\pi=.1$ (constant), $\rho=.6$ test for which the score distribution was almost singular. For closed intervals in the case of a skewed score distribution, it is difficult to make any recommendation based on the data at hand, unless the items are only moderately associated ($\rho \leq .3$). If this is the case, then the standard procedure will yield relative errors probably smaller than about ... When the items are more strongly associated, however, the resulting scor... become severely skewed and while the upper tail of the $\underline{V}$ distribu... ... nonulous, the situation in the lower tail is less predictable: for the three Ross tests the lower tail is too "light", for the $\pi=.2$ distribution it is about right and for the $\pi=.1$ distribution the procedure falls completely apart. Before summarizing, let us enumerate specifics of this investigation which necessarily limit the generalizations. They are:

1. Sample data from thirty respondents were simulated.
2. The number of items ranged from 10 to 30.
3. The normal ogive item characteristic curve related the trait being measured to the probability of a correct response.
4. Latent responses were sampled from multivariate normal distribution.
5. Tests simulated had a "single factor" structure.
6. For main body of results, latent item intercorrelations were constant.
7. Only 90% and 95% intervals were considered.

If a researcher has test data which has these characteristics, he may wish to consider the following recommendations:

1. For tests with item difficulty distributions which are widely spread about a median of .5, use the procedure but realize that it will tend to be conservative.
2. For tests with item difficulty distributions which are more homogeneous about a median difficulty of .5, use the procedure realizing that there will be a slight tendency for "too few" open intervals to cover the true parameter if the items are strongly associated.
3. For extremely skewed test score distributions the safe recommendation is to construct open intervals using the 97.5th percentile of the F distribution for nominal 95% intervals. The procedure will tend to be conservative.
4. For mildly skewed test score distributions no blanket recommendation is possible based on the data. However, if item intercorrelations are modest so that the resulting score distribution has an interior mode and $|\hat{r}.|$ is no more than about .4 or .5, the data suggest that the standard procedure will lead to relative errors of no more than 20% to 30%.

It is not surprising that in situations where the item difficulty is fairly homogeneous and different from .5 and the items are highly related that the usual robustness of the F distribution is not sufficient to provide serviceable inferences (The reader is referred to Mandeville (1969) for an extensive investigation related to hypothesis testing in repeated measures designs where the repeated measure is binary). However, in most of these cases where the approximation did not prove useful, true ρ_{20} was rather large (greater than .80). In situations where true ρ_{20} was less than .80, the parametric procedure did provide useful results. Since the concern of a researcher for the reliability of his measurements is usually inversel related to ρ_{20}, the practical value of these results appear great.

References

Baker, F.B., Empirical Determination of Sampling Distributions of Item Discrimination Indices and a Reliability Coefficient. Madison, Wisconsin. University of Wisconsin, 1962.

Cronbach, L.J. "Coefficient Alpha and the Internal Structure of Tests" Psychometrika, 1951, 16, 297-334.

Feldt, Leonard S. "The Approximate Sampling Distribution of Kuder-Richardson Reliability Coefficient Twenty" Psychometrika, 1965, 30, 357-370.

Feldt, Leonard S. "A Test of the Hypothesis That Cronbach's Alpha or Kuder-Richardson Twenty is the Same for Two Tests" Psychometrika, 1969, 34, 363-373.

Kendall, M.G. and Stuart A. The Advanced Theory of Statistics, Volume 2, Inference and Relationship. New York, Hafner, 1961.

Kuder, G.F. and Richardson, M.W. "The Theory of the Estimation of Test Reliability Psychometrika, 1937, 2, 151-160.

Lord, F.M. "Sampling Fluctuations Resulting from the Sampling of Test Items" Psychometrika, 1955, 20, 1-22.

Mandeville, G.K. "A Monte Carlo Investigation of the Adequacy of Standard Analysis of Variance Procedures for Dependent Binary Variates" Unpublished Ph.D Thesis, University of Minnesota, 1969.

Lord, F.M. and Novick, M.R. Statistical Theory of Mental Test Scores, Reading, Massachusetts. Addison-Wesley, 1968.

Nitko, Anthony J. and Feldt, Leonard S. "A Note on the Effect of Item Difficulty on the Sampling Distribution of KR_{20}" American Educational Research Journal, 1969 6, 433-437.

Payne, W.H. and Anderson, D.E. "Significance Levels for the Kuder-Richardson Twenty An Automated Sampling Experiment Approach" Educational and Psychological Measurement, 1968, 28, 23-39.

Ross, John "An Empirical Study of a Logistic Mental Test Model" Psychometrika, 1966, 31, 325-340.

Scheffe, Henry The Analysis of Variance, New York, Wiley, 1959.

TABLE 1

A Partial Comparison of Theoretical and Empirical 90th and 95th Percentiles of r_{20} Distributions Reported by Mitko and Feldt (1969)

Test ρ_{20}'s*		Theoretical 5th Percentile	Empirical 5th Percentile		Theoretical 10th Percentile	Empirical 10th Percentile	
I†	II		I	II		I	II
.554	.558	.356	.352	.364	.408	.411	.419
9	.690	.551	.559	.561	.586	.594	.594
.770	.771	.666	.671	.675	.693	.700	.700
.825	.826	.746	.755	.753	.766	.773	.772
.864	.865	.804	.810	.809	.820	.824	.824

*An average of the two ρ_{20} entries in a row was used in the computations to obtain the theoretical percentile.

†I and II refer to Mitko and Feldt's "Concentrated" and "Spread out" item difficulty distributions, respectively.

TABLE 2

Item Difficulties (π_i) for Three Ten Item Tests Simulated, Average Difficulties and Standard Deviations of the Item Difficulty Distributions

TEST	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	$\bar{\pi}$	σ_{π}
HET	.1	.2	.3	.4	.5	.5	.6	.7	.8	.9	.50	.245
HOM	.3	.35	.4	.45	.5	.5	.55	.6	.65	.7	.50	.122
HARD	.1	.15	.2	.25	.3	.3	.35	.4	.45	.5	.30	.122

TABLE 3

Descriptive Data and Chi-square Goodness of Fit Statistics for the 27 Tests Simulated

		HET					HOM					HARD				
ρ	k	ρ_{20}	$\bar{r}_{20}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	χ^2	ρ_{20}	$\bar{r}_{20}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	χ^2	ρ_{20}	$\bar{r}_{20}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\chi}^2$
	10	.36	.30	.01	-.22	5.8	.40	.36	-.02	-.42	13.9	.37	.33	.42	-.14	11.5
.10	20	.53	.50	-.01	-.25	28.5*	.57	.54	.02	-.29	4.8	.54	.51	.42	-.06	4.5
	30	.63	.60	.02	-.20	2.8	.66	.65	-.01	-.30	13.2	.64	.60	.37	-.11	11.0
	10	.65	.63	.01	-.59	8.5	.70	.67	.01	-.80	8.1	.67	.65	.61	-.25	5.5
.30	20	.79	.78	-.01	-.58	11.8	.82	.81	-.01	-.78	12.0	.80	.79	.59	-.26	8.0
	30	.85	.85	.01	-.61	23.2*	.87	.87	.01	-.77	20.3*	.86	.85	.64	-.19	17.6*
	10	.82	.82	-.01	-.96	51.5*	.86	.86	.00	-1.27	14.2	.85	.84	.75	-.53	12.0
.60	20	.91	.90	-.01	-.98	28.2*	.93	.92	-.03	-1.26	6.0	.92	.91	.76	-.49	21.3*
	30	.94	.93	.00	-.99	17.2*	.95	.95	-.02	-1.26	14.5	.94	.94	.77	-.48	17.6*

*$\chi_{.95} = 16.9$

The estimates of skewness and kurtosis of the total score distributions were computed using the sample moments obtained for all the scores generated in a computer run. For normal $\gamma_1 = 0$, $\gamma_2 = 0$. See Scheffe (1959), p. 331

TABLE 4

Empirical Probabilities of Incorrect Confidence Statements for Open-Ended and Closed Confidence Intervals on ρ_{20}.

			HET(.1<Π_i<.9)				HOM(.3<Π_i<.7)				HARD(.1<Π_i<.5)			
ρ	k	α	E_1	E_2	E_{2H}†	E_{2L}	E_1	E_2	E_{2H}	E_{2L}	E_1	E_2	E_{2H}	E_{2L}
.10	10	.10	095*	106	052	054	108	103	060	043	107	088	053	035
		.05	052	055	025	030	060	041	029	012	053	049	028	021
	20	.10	090	100	056	044	094	098	042	054	114	122	066	056
		.05	056	052	030	022	042	046	018	028	066	058	032	026
	30	.10	102	094	056	038	108	120	070	050	090	094	050	044
		.05	056	058	034	024	070	064	038	026	050	038	014	024
.30	10	.10	097	070	036	034	108	099	050	049	116	110	059	051
		.05	036	030	019	011	050	045	025	020	059	054	028	026
	20	.10	086	072	036	036	120	092	048	044	112	114	064	050
		.05	036	036	018	018	048	034	016	018	064	048	018	030
	30	.10	106	070	044	026	084	072	048	024	140	134	084	050
		.05	044	032	018	014	048	038	024	014	084	066	042	024
.60	10	.10	065	048	028	020	123	091	059	032	126	133	070	063
		.05	028	019	013	006	059	048	038	010	070	079	043	036
	20	.10	084	076	048	028	118	100	060	040	120	122	072	050
		.05	048	030	016	014	060	046	030	016	072	078	046	032
	30	.10	106	066	038	028	116	066	038	028	146	128	078	050
		.05	038	026	012	014	038	028	018	010	078	076	044	032

* Decimals omitted in body

† A reversal of the implication of statements on page 5 has been made for mnemonic reasons so the E_{2H} is the proportion of times that the total interval was "too high", i.e., $C_{2L}>\rho_{20}$. Similarly E_{2L} indicates the proportion of times that the interval was "too low", i.e., $C_{2H}<\rho_{20}$.

TABLE 5

Test characteristics and simulation results for tests W, X, Y, and Z

Test characteristics include number of items (k), average difficulty ($\bar{\pi}$), standard deviation of the difficulty distribution (σ_{π}), average item intercorrelation $\bar{\rho}_{ij}$ and population ρ_{20}. The average sample reliability estimates ($\bar{r}_{20}$) and skewness and kurtosis of the score distributions ($\hat{\gamma}_1$ and $\hat{\gamma}_2$) are also reported. At the right are empirical probabilities of incorrect open and closed, nominal 90% and 95% confidence intervals.

Test	k	$\bar{\pi}$	σ_{π}	$\bar{\rho}_{ij}$	ρ_{20}	$\bar{r}_{20}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	α	E_1	E_2	E_{2U}	E_{2L}
W	12	.62	.053	.67	.910	.91	-.485	-1.18	.10	142*	138	096	042
									.05	096	080	061	019
X	16	.54	.036	.76	.95	.95	-.20	-1.53	.10	161	130	104	026
									.05	104	077	068	009
Y	13	.49	.094	.70	.920	.92	.05	-1.43	.10	117	081	058	023
									.05	058	039	029	010
Z	18	.58	.120	.65	.93	.93	-.35	-1.21	.10	105	084	057	027
									.05	057	041	033	008

* Decimals omitted in body.