JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 16, NO. 2 SUMMER 1979

AN EMPIRICAL COMPARISON OF COEFFICIENT ALPHA, GUTTMAN'S LAMBDA - 2, AND MSPLIT MAXIMIZED SPLIT-HALF RELIABILITY ESTIMATES

JOHN C. CALLENDER
Shell Oil Company
H. G. OSBURN
University of Houston

Guttman (1945) derived six different types of coefficients and showed that each was a lower bound to reliability defined as the ratio of true score variance to observed score variance. Guttman's formulas for the lower bounds (in population notation) are given below:

$$\lambda_1 = 1 - \frac{\sum_{j=1}^{n} \sigma_j^2}{\sigma_x^2} \tag{1}$$

$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{n}{n-1}\Gamma_2}}{\sigma_x^2} \tag{2}$$

$$\lambda_3 = \frac{n}{n-1}\lambda_1 \tag{3}$$

$$\lambda_4 = 2\left[1 - \frac{\sigma_a^2 + \sigma_b^2}{\sigma_x^2}\right] \tag{4}$$

$$\lambda_5 = \lambda_1 + \frac{2\sqrt{\bar{\Gamma}_2}}{\sigma_x^2} \tag{5}$$

$$\lambda_6 = 1 - \frac{\sum_{j=1}^{n} e_j^2}{\sigma_x^2} \tag{6}$$

In the formulas σ_x^2 is the variance of the observed scores on the composite measure; σ_j^2 is the variance of a single item j; Γ_2 is the sum of the squares of the covariances between items, a sum which includes $n(n-1)$ terms; σ_a^2 is the variance of observed scores from one part or "half" of the composite and σ_b^2 is the variance of the remaining part or half of the composite; $\bar{\Gamma}_2$ is the sum of the squares of the covariances of one particular item with the remaining $n-1$ items, for whichever item gives the largest such sum; e_j^2 is the variance of the errors of estimate of item j from its linear multiple regression on the remaining $n-1$ items.

Guttman showed that:

$$\lambda_1 < \lambda_3 \leq \lambda_2 \tag{7}$$

Since λ_1 is always less than both λ_2 and λ_3, it is also always an underestimate of reliability. Since λ_2 is sometimes greater than λ_3, it follows that it could be a more accurate reliability lower bound than λ_3, but would never be less accurate than λ_3.

The authors gratefully acknowledge the support of C. Paul Sparks, Exxon Co., USA, who provided the computer and data resources for this research, and thank Jack Greener for reviewing an earlier draft.

The λ_4 coefficient requires special discussion because, although there is only a single formula, it defines a whole family of coefficients. Each is a lower bound to ρ^2_{XT}, but can be defined by a different allocation of the n items to part "a" and part "b". The common interpretation of λ_4 is that it would be a split-half reliability coefficient, such as could be defined by putting n/2 items in part "a" and the remaining n/2 items in part "b." But it is not necessary that the two parts have the same number of items in order for λ_4 to be a lower bound to ρ^2_{XT}. Guttman pointed out that no assumption of parallelism or equivalence of the two parts was necessary in order for *any* λ_4 to be a lower bound. Thus it is not necessary for the two parts to have the same variance.

Guttman concluded that λ_4 would be a particularly useful lower bound because some splits could be larger than others. By finding a split with a larger λ_4, reliability would be more accurately "lower bounded." He commented that it was frequently easy to find λ_4 splits which gave better lower bounds than λ_3 or λ_2. The best internal consistency reliability estimate would be found by computing every possible lower bound, including every possible split coefficient, and using the largest. Computational effort would seem to be the only limiting factor, provided one has a population variance/covariance matrix.

The formula for Guttman's λ_3 lower bound is the same as that for two better known reliability coefficients, namely the Kuder-Richardson formula 20 and Cronbach's Coefficient Alpha. Cronbach (1951) demonstrated that there was a relationship between Coefficient Alpha and the possible split-half coefficients. Coefficient Alpha would be equal to the mean of all possible split-half coefficients. In Guttman's notation we would have that:

$$\lambda_3 = E(\lambda_4) \tag{8}$$

Novick and Lewis (1967) derived the conditions under which λ_4 and λ_3 would actually be equal to reliability. The necessary and sufficient condition for:

$$\lambda_3 = \lambda_4 = \lambda_2 = \rho^2_{XT} \tag{9}$$

is that all items must be "essentially tau-equivalent." This means that every person's true score on one item must differ from his true score on another item by an additive constant. It implies some important constraints on the true score, error score, and observed score covariance matrices. The entries in the item true score variance/covariance matrix must all be equal, since all item true scores would be perfectly correlated. The item error variances, however, need *not* all be equal. The item error covariances would all be assumed to be zero, following the usual classical assumptions. Thus the observed score variance/covariance matrix has possibly differing variance elements, but all covariances must be equal. Item means could also differ from one another. Thus, the circumstance in which there would be some λ_4 greater than λ_3 is when the population observed item covariances are not all equal. This would seem to be a common state of affairs.

Jackson and Agunwamba (1977) have given a mathematical analysis of Guttman's lower bounds, λ_1 through λ_6, in terms of the population covariance matrix properties that would lead to one coefficient being larger than another. With regard to the λ_4 split-half coefficient, only one of the possible split-halves was considered of interest, namely the one which would give the largest λ_4 value. In treating the λ_4 coefficient, they did not require that the two halves have an equal number of items. Maximum λ_4 was the largest from among the 2^{n-1} possible allocations of items into two parts. A new greatest lower bound (g. 1. b.) was developed and shown to be a lower bound to reliability which would be as

large as the greatest of all of Guttman's lower bounds, including the largest λ_4. The analysis of the differences in lower bounds as related to covariance matrix properties led Jackson and Agunwamba to conclude that unless the population covariance matrix has very special properties, unlikely to be encountered in practice, maximum λ_4 is the only one which may actually be the *greatest* lower bound. There is no simple formula for finding the g. l. b., rather it must be found by a computer search procedure which was described in a companion paper by Woodhouse and Jackson (1977). After computing the g. l. b. and other lower bounds on actual test data, Woodhouse and Jackson concluded that the maximum split-half λ_4 was a considerable improvement over λ_3 and λ_2, while the gain in going to the g. l. b. from maximum λ_4 was fairly modest. It was also mentioned that for numbers of subtests (parts) up to 17, less time was required to compute the largest λ_4 than to compute the g. l. b. A caution was given about using the bounds safely with sample covariance matrices computed from the scores of a modest number of persons, because of the lack of knowledge of sampling characteristics. The authors seemed to doubt that the sampling distributions would be easily obtained mathematically and indicated that simulation would be necessary to study sampling bias and variance.

Callender and Osburn (1977a) have reported a method, called MSPLIT, to determine a large split-half L_4 without having to compute every possible split coefficient. This is increasingly advantageous computationally as the number of items increases. When applied to actual test data with 10 items per test, it was found that on the average the MSPLIT L_4 was only about .02 smaller than the largest split-half L_4. The MSPLIT L_4's were considerably larger than the L_3 and odd-even split L_4 coefficients.

THE SAMPLING PROBLEM FOR MAXIMIZED COEFFICIENTS

The problems that arise when sample values are maximized to estimate population parameters are well known and typically result in overestimation. One would expect therefore that sample values of the g. l. b., maximized L_4's, and L_6 would tend to overestimate their respective population counterparts. Further, if that population counterpart itself is a "good" lower bound, it may be that actual population reliability could be overestimated by the maximized sample coefficient. Maximized population coefficients do not have this problem because there is no sampling variance to confound. Obviously there is a real problem which must be overcome in order for the maximized coefficients to find application on anything other than giant samples which could be regarded as "the population." There are two issues which might be dealt with: the "bias" issue, or how can a sample coefficient be adjusted in order to correctly estimate its population counterpart, and the "accuracy" issue, or how can sample coefficients be treated in order to correctly estimate the population reliability or the greatest population lower bound?

EMPIRICAL STUDY OF MSPLIT COEFFICIENTS: OBJECTIVES AND METHOD

An empirical study of MSPLIT maximized split-half L_4's was undertaken to assess the degree of overestimation due to capitalization on sampling error and to investigate methods of handling it. Another issue of interest was to determine *how much* better accuracy might result from using MSPLIT coefficients instead of L_2 or L_3. The latter

Table 1

Population and Sample Reliability Coefficients for 10-Item Tests

	Population Coefficient (N=380)	Mean of 10 Total Sample Coefficients (N=100)	Mean of 20 Original Sample Coefficients (N=50)	Mean of 20 Holdout Sample Coefficients (N=50)
TEST 10-1				
Largest L_4	.235	.367	.467	
MSPLIT L_4	.222	.341	.424	.028
L_2	.151	.182	.209	
L_3	.127	.093	.080	
TEST 10-2				
Largest L_4	.279	.391	.494	
MSPLIT L_4	.279	.363	.449	.109
L_2	.174	.203	.231	
L_3	.146	.120	.111	
TEST 10-3				
Largest L_4	.381	.501	.575	
MSPLIT L_4	.337	.474	.554	.225
L_2	.286	.335	.362	
L_3	.261	.270	.263	
TEST 10-4				
Largest L_4	.478	.615	.673	
MSPLIT L_4	.478	.595	.646	.445
L_2	.390	.463	.479	
L_3	.372	.417	.414	
TEST 10-5				
Largest L_4	.541	.590	.658	
MSPLIT L_4	.541	.564	.601	.401
L_2	.426	.415	.421	
L_3	.412	.365	.342	
TEST 10-6				
Largest L_4	.661	.716	.749	
MSPLIT L_4	.652	.700	.723	.571
L_2	.591	.575	.579	
L_3	.587	.547	.532	
TEST 10-7				
Largest L_4	.709	.748	.789	
MSPLIT L_4	.705	.735	.774	.648
L_2	.655	.652	.659	
L_3	.651	.638	.634	
TEST 10-8				
Largest L_4	.724	.775	.802	
MSPLIT L_4	.697	.764	.772	.683
L_2	.673	.675	.674	
L_3	.669	.661	.650	
TEST 10-9				
Largest L_4	.759	.795	.825	
MSPLIT L_4	.752	.775	.805	.691
L_2	.707	.684	.689	
L_3	.699	.661	.654	

Table 2

Population and Sample Reliability Coefficients for 40-Item Tests

	Population Coefficient (N=380)	Mean of 10 Total Sample Coefficients (N=100)	Mean of 20 Original Sample Coefficients (N=50)	Mean of 20 Holdout Sample Coefficients (N=50)
TEST 40-1				
MSPLIT L_4	.672	.802	.848	.600
L_2	.550	.578	.591	
L_3	.531	.534	.527	
TEST 40-2				
MSPLIT L_4	.834	.890	.917	.780
L_2	.767	.769	.774	
L_3	.762	.752	.746	
TEST 40-3				
MSPLIT L_4	.905	.943	.956	.888
L_2	.866	.870	.872	
L_3	.864	.864	.861	

in Table 1 it will be noted that the MSPLIT and largest L_4 coefficients of the total samples exceed those of the population, and those of the original samples further exceed those of the total samples. A reduction of the MSPLIT coefficients in the holdout samples is apparent. The conclusions based on Table 1 are also applicable to Table 2, except that the largest split-half coefficients were not obtained.

At first look it was surprising to find that the shrinkage of the MSPLIT L_4's brought them so far down as to be comparable to the L_3 coefficients. This is like falling from the top of the distribution of L_4's down to about the middle. Why should this occur? A little reflection on what was happening as a result of the maximization process provided a rationale for better use of the original and holdout sample MSPLIT L_4's. Suppose that in the population there is a largest split-half λ_4 and we happen to know what allocation of items would produce it. If we had this knowledge and a sample of data for computing reliability, we should allocate the items in the same manner as in the population in computing the sample L_4. We might very well expect such a coefficient to be a relatively unbiased estimator of that largest λ_4 across several samples. In reality, we do not know what allocation of items this would be, so we divide our sample and apply MSPLIT to one subsample. This does two things for (or to) us. First, because we have added random sampling error to the covariances, we find that MSPLIT maximizes some of it and gives us an L_4 coefficient that is larger than what we would have obtained if we had used the "correct" allocation of items from the population. Second, we have come to the wrong conclusion about which allocation to use. Now we apply that "wrong" allocation of items to the other subsample. There is no addition of sampling variation to the cross-validated L_4 value in that sample because no maximizing was done on it. Furthermore, since we are using the "wrong" allocation of items, we find that the L_4 coefficient is actually less than the L_4 we would have obtained if we had used the correct allocation of items.

Thus we find that original sample MSPLIT L_4's tend to be biased high relative to the L_4 for the correct population maximized λ_4, and that holdout sample MSPLIT L_4's are biased low relative to it. It seemed plausible that a better estimator of the largest

population λ_4 might be obtained by averaging the four MSPLIT L_4 coefficients, where two are maximized original samples and the other two are holdout sample coefficients. The resulting mean coefficient will be referred to as the "MSPLIT estimator" of reliability. The issue we now turn to is that of the accuracy of the MSPLIT estimator, total sample L_2, and total sample L_3 as competing sample reliability estimators.

Because the covariance matrices were not simulated to produce a known true reliability, we cannot compare the various sample coefficients with it as a measure of accuracy. However, we can make an analysis of accuracy relative to the greatest computed population lower bound. A coefficient which more accurately estimates that value would also be a more accurate estimator of the population true reliability. In the case of the 10-item tests we use the largest population λ_4 as the standard. For the 40-item tests, the standard is the MSPLIT population λ_4.

In studying the bias or sampling variation of statistical indices, the ideal situation is to have a very large number of samples, each of relatively small size. But how many samples are needed? We found that the results were very consistent after drawing just 10 samples. As a check on how well the sample results might be agreeing with what would be expected from an infinite number of samples, we compared the mean of the sample L_3 coefficients with their expected values based on the population λ_3. Table 3 shows that the mean L_3 coefficients were within $-.058$ to $+.035$ of the expected value after 10 samples. The expected value of the L_2 coefficients and the MSPLIT estimator could not be computed directly, since no formula yet exists for this purpose. However, it was possible to make an empirically based projection of their expected value. For each test, the relative positions of the MSPLIT estimator, L_2, and L_3 were remarkably constant across samples. If one went up, so did the other two. As a check on this consistency it was found that the regression of the MSPLIT estimator on L_3 and the regression of L_2 on L_3 across samples were statistically significant at .05 level or better for every test. This provided a way of projecting the expected value of the MSPLIT estimator and L_2 coefficient for each test. It is reasonable to conclude that, if the L_3's averaged out to something less than

Table 3

Comparison of Observed Mean and Expected Value of Total Sample L_3 Coefficients

Test	Population λ_3	Expected Value of Total Sample L_3's	Mean of 10 Total Sample L_3's	Difference Between Expected Value and Mean of Total Sample L_3's
10-1	.127	.110	.093	+.017
10-2	.146	.129	.120	+.009
10-3	.261	.246	.270	-.024
10-4	.372	.359	.417	-.058
10-5	.412	.400	.365	+.035
10-6	.587	.579	.547	+.032
10-7	.651	.644	.638	+.006
10-8	.669	.662	.661	+.001
10-9	.699	.693	.661	+.032
40-1	.531	.522	.534	-.012
40-2	.762	.757	.752	+.005
40-3	.864	.861	.864	-.003

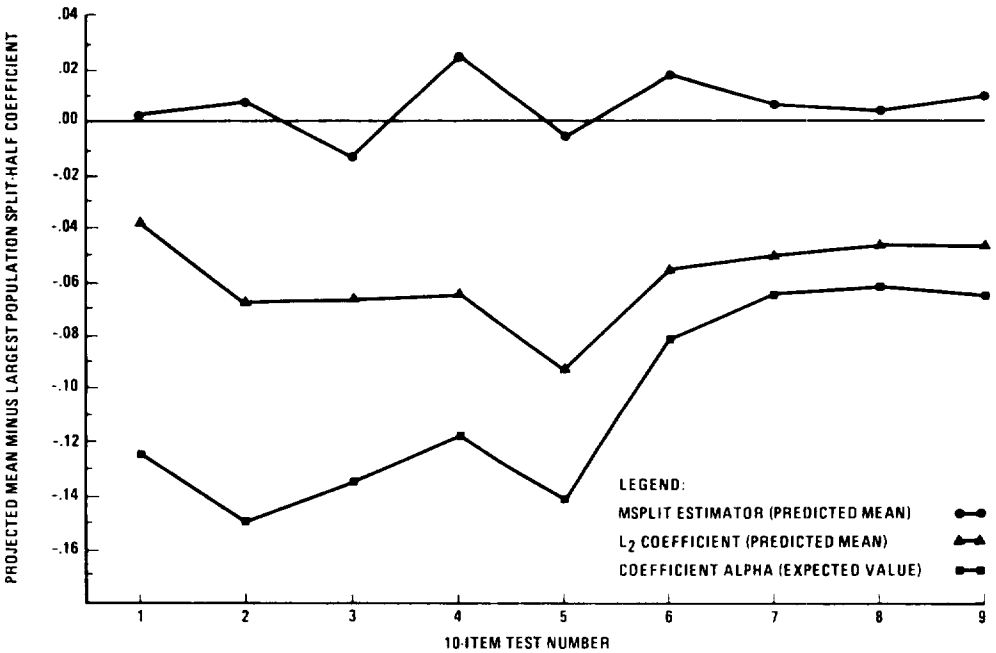

Figure 1. Difference between long-run projected mean of total sample estimators and largest population split-half coefficient in 10-item tests.

their expected value due to chance sampling error, the MSPLIT estimator and L_2 coefficients were also lower than their expected value. By simply substituting the expected value of L_3 into the regressions of the MSPLIT estimator and L_2 on L_3, we obtained a projection of the expected value of those coefficients. In going from results based on 10 samples to projected results based on an infinite number of samples, it must be recognized that only small adjustments were made. This is so because the adjustments were based on the small differences between observed mean L_3's and their expected value reported in Table 3. It must also be recognized that relative differences among the three kinds of coefficients were maintained by the adjustment procedure.

The difference between the long-run projected mean of each type of sample estimator and the largest population split-half coefficient is shown in Figure 1 for the 10-item tests. Figure 2 provides similar information regarding the 40-item tests, except that the difference is computed between the long-run expected means and the MSPLIT population split-half coefficient.

Figures 1 and 2 show that the largest split-half coefficient (and hence the population reliability) is more accurately estimated by computing the MSPLIT estimator than by computing sample L_2 or L_3 coefficients. This conclusion is supported for either a large or a small number of items and for a wide range of test reliability. For 40-item tests, the MSPLIT population coefficient was consistently overestimated by a small amount. Of course, this does not necessarily imply that the population reliability was overestimated, since the MSPLIT population coefficient may not have been the largest λ_4. The inaccuracy of L_3 (Coefficient Alpha) is rather striking, with reliability underestimated by more than .1 for several tests. As L_3 increases in Figures 1 and 2, there is a tendency

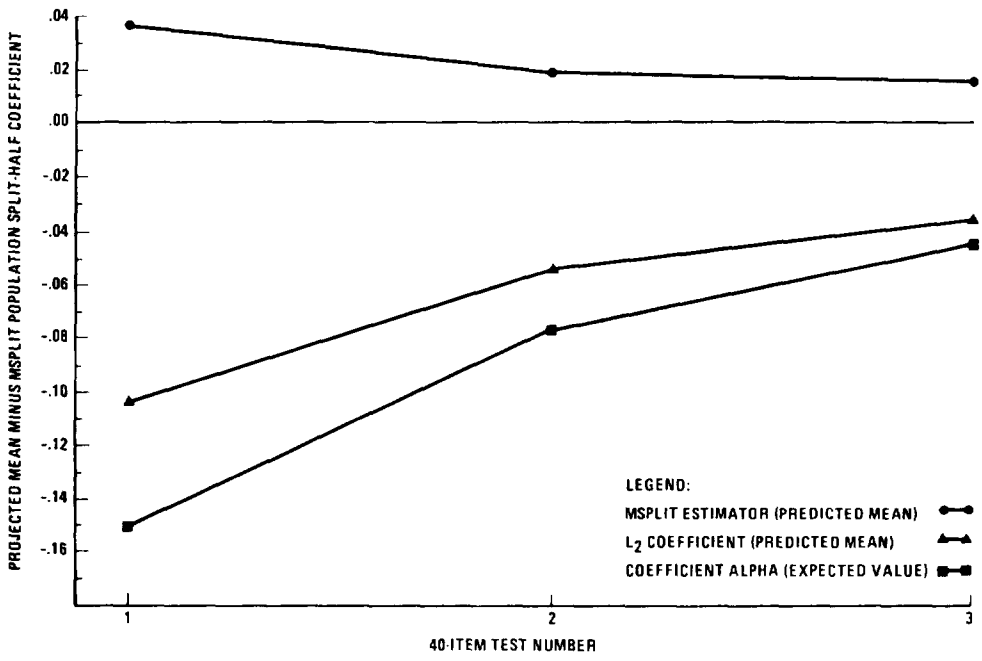

Figure 2. Difference between long-run projected mean of total sample estimators and MSPLIT population split-half coefficient in 40-item tests.

for the discrepancy between it and the maximized population split-half coefficient to be reduced. This may result from the fact that as L_3 (the average split-half coefficient) approaches 1.0, there is a necessary reduction in the possible discrepancy between L_3 and the largest split-half coefficient.

The standard deviation over the 10 samples of L_3 (Coefficient Alpha), Guttman's L_2, and the MSPLIT estimator is shown in Figure 3 for each test. The sampling variation of the three estimators is similar, except that the MSPLIT estimator was slightly less variable in the 40-item tests. The degree of sampling variation in the low reliability tests was rather great, indicating that with a sample size of only $N = 100$, any of the sample estimators has a high probability of missing the population reliability by a large amount.

CONCLUSIONS

The MSPLIT estimator was clearly a better estimator of reliability than the usual L_3 or L_2. It effectively overcame the problem of capitalization on sampling variation, even on very small sample sizes. With these data, the absolute magnitude of the underestimation of reliability by L_3 was so large as to give one pause over the kind of error that could result from having used it. For example, substantial amounts of *over*correction for attenuation could have resulted if L_3 had been used instead of the MSPLIT estimator. Internal consistency reliabilities may also be used in order to infer the power of an experiment or of a statistic. Use of an underestimate of reliability would result in an underestimation of power.

We think that there are several additional kinds of research that would be interesting

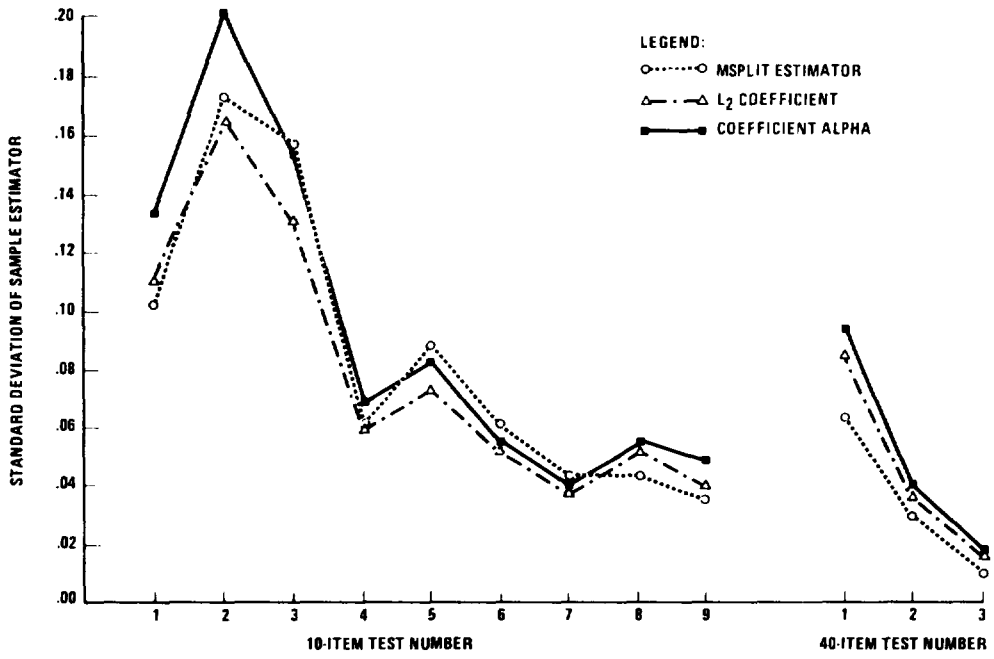

Figure 3. Efficiency of sample estimators.

and informative. One comparison would be to simulate population covariance structures of known reliability that are not essentially tau-equivalent. Sample MSPLIT estimators could then be compared with the true reliability. We already know that the MSPLIT estimator is more accurate than L_2 and L_3 coefficients relative to true reliability, but we don't know the absolute magnitude of the difference between the MSPLIT estimator and true reliability.

Simulation studies could also be used to explore the sampling distribution of MSPLIT estimators relative to other reliability coefficients on non-essentially tau-equivalent data. The present study did not contain a sufficient number of samples to draw firm conclusions on this matter.

Finally, we note that a program for computing and cross-validating MSPLIT coefficients is available (Callender & Osburn, 1977b).

REFERENCES

CALLENDER, J. C., & OSBURN, H. G. A method for maximizing split-half reliability coefficients. *Educational and Psychological Measurement*, 1977, **37**, 819–826. (a)

CALLENDER, J. C., & OSBURN, H. G. A computer program for maximizing and cross-validating split-half reliability coefficients. *Educational and Psychological Measurement*, 1977, **37**, 787–790. (b)

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297–334.

GUTTMAN, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, **10**, 255–282.

JACKSON, P. H., & AGUNWAMBA, C. C. Lower bounds for the reliability of the total score on

a test composed of non-homogeneous items: Algebraic lower bounds. *Psychometrika,* 1977, **42,** 567–578.

NOVICK, M. R., & LEWIS, C. Coefficient alpha and the reliability of composite measurements. *Psychometrika,* 1967, **32,** 1–13.

WOODHOUSE, B., & JACKSON, P. H. Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: A search procedure to locate the greatest lower bound. *Psychometrika,* 1977, **42,** 579–591.

AUTHORS

CALLENDER, JOHN C. *Address:* Shell Oil Company, One Shell Plaza, Box 2463, Houston, Texas 77001. *Title:* Senior Employee Relations Analyst. *Degrees:* B.A. Butler University, Ph.D. University of Houston. *Specialization:* Industrial and Organizational Psychology.

OSBURN, H. G. *Address:* Department of Psychology, University of Houston, Houston, Texas 77004. *Title:* Professor of Psychology. *Degrees:* B.S., Ph.D. University of Michigan. *Specialization:* Industrial and Organizational Psychology.