Estimating Measures of Pass-Fail Reliability From Parallel Half-Tests

David J. Woodruff and Richard L. Sawyer
American College Testing Program

Two methods are derived for estimating measures of pass-fail reliability. The methods require only a single test administration and are computationally simple. Both are based on the Spearman-Brown formula for estimating stepped-up reliability. The non-distributional method requires only that the test be divisible into parallel half-tests; the normal method makes the additional assumption of normally distributed test scores. Bias for the two procedures is investigated by simulation. For nearly normal test score distributions, the normal method performed slightly better than the non-distributional method, but for moderately to severely skewed or symmetric platykurtic test score distributions the non-distributional method was superior. Test results from a licensure examination are used to illustrate the methods. *Index terms: Cohen's kappa, licensure examinations, pass-fail reliability, reliability, Spearman-Brown formula.*

A primary component of the *Standards for Educational and Psychological Testing* (American Psychological Association, 1985) concerning licensure and certification examinations requires test publishers to report the reliability of pass-fail decisions (PF reliability). Hambleton and Novick (1973) proposed θ, the proportion of consistently classified examinees, as a measure of PF reliability. Swaminathan, Hambleton, and Algina (1974) suggested that Cohen's (1960) kappa coefficient, κ, should replace θ. Coefficient κ is the proportion of consistently classified examinees, corrected for chance. Though it commonly is used to measure association rather than agreement, ϕ, the Pearson correlation between two dichotomous variables, equals κ under certain circumstances. Thus, ϕ may also be used as a measure of PF reliability. Fleiss (1981) critically discussed θ and κ as well as other indices of agreement.

If two parallel test forms are available for administration to the same sample of examinees, then estimates for θ and κ are easily obtained by the method of moments (MOM). If only one form of the test may be administered, then obtaining estimates for θ and κ becomes much more difficult, both theoretically and computationally. Huynh (1976) developed a procedure for estimating θ and κ based on a beta-binomial model that requires only one test administration. The computations involved are quite tedious, but Huynh (1976) also suggested a simpler method based on a normal approximation. Peng and Subkoviak (1980) further simplified Huynh's (1976) approximate method and presented evidence suggesting that their simplified procedure is superior to Huynh's. Brennan (1981) and Subkoviak (1988) supplied tables which make the computations for Peng and Subkoviak's (1980) procedure relatively simple. Subkoviak (1984) discussed several other methods for estimating θ and κ when only one test form is available.

This paper presents two theoretically and computationally simple methods to estimate both θ and

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 1, March 1989, pp. 33–43
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/010033-11$1.80

κ from a single test administration when the test is divisible into parallel half-tests. One method is based on normal theory; the other makes only minimal distributional assumptions. Bias for the two procedures is analyzed by using simulation techniques under a variety of test score distributions and test reliabilities.

Derivation of the Methods

Let X denote the total test and Y1 and Y2 the parallel half-tests (Lord & Novick, 1968) into which X is divisible. As will be shown later, the statistical assumptions defining parallelism for tests Y1 and Y2 may be relaxed so long as Y1 and Y2 are parallel (homogeneous) in content. Let A denote the dichotomous variable that equals 0 when an examinee fails X and equals 1 when an examinee passes X. The dichotomous variables B_1 and B_2 are similarly defined for Y1 and Y2. The three variables A, B_1, and B_2 require that passing scores be set for X, Y1, and Y2. The passing score for X is usually determined, at least in part, from criterion information distinct from the pass rate. However, the passing scores for Y1 and Y2 are set so that the pass rates for Y1 and Y2 are identical to that of X.

The proportion parameters describing the variables B_1 and B_2 may be expressed in the usual format of a 2 × 2 table as in Table 1. In Table 1, π_{00} is the proportion of examinees in the population of interest who fail both Y1 and Y2, and π_{01}, π_{10}, and π_{11} are defined analogously. Because the pass rate p is the same for B_1 and B_2, $\pi_{01} = \pi_{10}$. In terms of these proportion parameters,

$$\phi = 1-\pi_{01}/pq \quad , \qquad (1)$$

Table 1
Probabilities of Outcomes Associated With Parallel Half-Tests

	B_2		
B_1	0	1	Total
0	π_{00}	π_{01}	q
1	π_{10}	π_{11}	p
Total	q	p	1

$$\theta = \pi_{00}+\pi_{11} \quad , \qquad (2)$$

and

$$\kappa = (\theta-\theta_l)/(1-\theta_l) \quad , \qquad (3)$$

where $\theta_l = p^2 + q^2$. The parameter θ_l is the value of θ when B_1 and B_2 are statistically independent. Assuming that p is the same for B_1 and B_2, then $\kappa = \phi$, as Cohen (1960) first noted. He also stated that κ and ϕ are nearly identical if the pass rates differ by no more than .10.

Estimates for ϕ, θ, and κ obtained by substituting observed proportions into Equations 1, 2, and 3 would provide PF reliabilities for the half-tests Y1 and Y2. However, PF reliabilities are not obtained for the whole test X, which is what is desired.

Let Y1* and Y2* be the doubled-in-length versions of Y1 and Y2, with lengthening done according to the model of parallel measurements. Thus, Y1* and Y2* are parallel forms of X. Let B_1^* and B_2^* be the dichotomization of Y1* and Y2*, assuming that the passing scores for Y1* and Y2* are selected so that the pass rates are the same as for Y1, Y2, and X. Finally, let $\phi^* = \kappa^*$ and θ^* be the PF reliability coefficients corresponding to B_1^* and B_2^* (and, consequently, to A).

A Normal Theory Method

Modifying the Huynh-Peng-Subkoviak (HPS) procedure to fit the present model of parallel half-tests provides a simple way to estimate ϕ^* and θ^*. The HPS beta-binomial model assumption concerning item sampling can be dropped, but the bivariate normal approximation for test scores can be retained. Let $K_q = (K - \mu_x)/\sigma_x$, where K is the passing score on the total test, and μ_x and σ_x are the population mean and standard deviation for the total test. Under the model's assumptions, $q = P(Z \leq K_q)$, where Z is a standard normal random variable, and $p = 1 - q$. Furthermore,

$$\pi_{00}^* = P(Z_1{\leq}K_q,\, Z_2{\leq}K_q;\, \rho^*) \quad , \qquad (4)$$

where Z_1 and Z_2 have a standard bivariate normal distribution with correlation coefficient ρ^*.

To estimate K_q, both μ_x and σ_x can be replaced with their corresponding sample estimators. From the parallel half-scores on Y1 and Y2, the half-test reliability, ρ, can be estimated and then stepped up to an estimate, r_{SB}, of the full-test reliability, ρ^*, using the Spearman-Brown formula. The estimates for K_q and ρ^* can then be substituted in Equation 4 to estimate π^*_{00}. Values for the estimate of π^*_{00} can be found in Brennan's (1981) tables or in other tables of the bivariate normal distribution function. Estimates for the probabilities π^*_{01} and π^*_{11} can then be computed from the relationships $\pi^*_{00} + \pi^*_{01} = q$ and $\pi^*_{11} + \pi^*_{01} = p$. Because $\theta^* = \pi^*_{00} + \pi^*_{11}$ and $\phi^* = \kappa^* = 1 - \pi^*_{01}/pq$, the normal model estimates for π^*_{00}, π^*_{11}, and π^*_{01} can be immediately converted to estimates for θ^* and ϕ^*.

The practical difference between the method proposed here and the HPS method is the use of the stepped-up reliability estimate, r_{SB}, in place of KR-21. The basic theoretical difference is that the method proposed here relies on a parallel half-test model rather than on a beta-binomial model for the full test.

A Non-Distributional Method

In the normal-theory model above, the half-length reliability, ρ, was stepped up to the full-length reliability, ρ^*, by the Spearman-Brown formula. Alternatively, the half-length PF reliability, ϕ, could be stepped up directly: $\phi^*_{SB} = 2\phi/(1 + \phi)$. Substituting the expression for ϕ given in Equation 1 into this expression for ϕ^*_{SB} and simplifying yields

$$\phi^*_{SB} = 1 - \frac{\pi_{01}}{2pq - \pi_{01}} \quad , \tag{5}$$

where $p = \pi_{01} + \pi_{11}$ is the pass rate and $q = 1 - p$. Because $\phi^* = 1 - \pi^*_{01}/pq$, it follows that

$$\phi^*_{SB} - \phi^* = \frac{\pi^*_{01}}{pq} - \frac{\pi_{01}}{2pq - \pi_{01}} \quad . \tag{6}$$

Note that the pass rate $p = \pi_{01} + \pi_{11} = \pi^*_{01} + \pi^*_{11}$ is the same for both the half-length and the full-length tests. For non-zero π^*_{11}, the left-side difference in Equation 6 is 0 if and only if $\pi^*_{01} = [1/(1 + \phi)]\pi_{01}$. However, because $0 \leq \pi^*_{01} \leq \pi_{01}$, the first term of the right-side difference is greater than 0 and less than π_{01}/pq, leading to the following upper and lower bounds for the left-side difference in Equation 6:

$$-\frac{\pi_{01}}{2pq - \pi_{01}} \leq \phi^*_{SB} - \phi^* \leq \frac{\pi_{01}}{pq} - \frac{\pi_{01}}{2pq - \pi_{01}} \quad . \tag{7}$$

Because each side of this inequality approaches 0 as π_{01} approaches 0, ϕ^*_{SB} becomes a better approximation to ϕ^* as the half-test reliability increases, though ϕ^*_{SB} can be a useful approximation to ϕ^* when the half-test reliability is only moderate.

Technically, ϕ^*_{SB} is the reliability of a test composed of two parts with dichotomous scores B_1 and B_2 or, equivalently, the correlation between two parallel forms of such a test. These test scores would be trichotomous variables taking the values 0, 1, and 2. Consequently, the interpretation of ϕ^*_{SB} as the correlation between B^*_1 and B^*_2 is an approximation, because B^*_1 and B^*_2 are dichotomous variables.

More specifically, let $C = B_1 + B_2$ and $C' = B'_1 + B'_2$ where the prime denotes a parallel measurement. The coefficient ϕ^*_{SB} equals the correlation between the two parallel variables, C and C', both of which may take the values of 0, 1, or 2. For the variable C, the value 0 occurs if both B_1 and B_2 equal 0. The value 1 occurs if the examinee passes one half-test but fails the other. If both B_1 and B_2 equal 1, then C takes the value of 2. The values of C' are similarly defined in terms of B'_1 and B'_2.

If B_1 and B_2 are derived from reasonably reliable tests, then relatively few examinees should have values of 1 on C and C'. Assume that the group of examinees with 1 on C is approximately the same as the group of examinees with 1 on C'. Let the dichotomous variables D and D' be defined by dividing this group of examinees in half and assigning one half scores of 0 and the other half scores of 2.

The dichotomous variables D and D' will have approximately the same means as C and C', but their variances will be larger. The covariance between D and D' will also be larger than the covariance between C and C'. Consequently, the correlation between D and D' should be approximately equal to the correlation ϕ^*_{SB} between C and C'. However, because D and D' may be seen as a dichotomous grouping of the trichotomous variables C and C', ϕ^*_{SB} may be close to but slightly greater than the correlation between D and D' because of attenuation due to grouping. The variables D and D' take the values of 0 or 2 while the variables B_1^* and B_2^* take the values of 0 or 1. Although the variables D and D' are defined differently than the variables B_1^* and B_2^*, their values are linearly related with a slope equal to 2 and an intercept equal to 0. Because the variables D and D' represent the variables B_1^* and B_2^* on a different scale, ϕ_{SB}^* may be interpreted as an approximation to ϕ^*, the correlation between B_1^* and B_2^*.

The approximation ϕ_{SB}^* is related to θ^* as follows. First, algebraic manipulations show that $\pi_{11}^* = pq\phi^* + p^2$ and $\pi_{00}^* = pq\phi^* + q^2$. Combining these equations with the relationship $\theta^* = \pi_{00}^* + \pi_{11}^*$ results in

$$\theta_{SB}^* = 2pq\phi_{SB}^* + p^2 + q^2 \quad . \tag{8}$$

The relationship to κ^* is simpler: $\kappa_{SB}^* = \phi_{SB}^*$.

In practice, constructing exactly parallel half-tests from a single test may not be possible, and thus π_{01}^* may not equal π_{10}^*. If the difference in the observed proportions p_{01} and p_{10} corresponding to π_{01} and π_{10} is minor in relation to the sample size, then p_{01} and p_{10} may be replaced by their average, and the marginal proportions adjusted accordingly. The above formulas are then applied to the modified 2 × 2 table of observed proportions.

Simulation Study

Method

A monte carlo investigation was conducted to evaluate the accuracy of the non-distributional and normal procedures. An important application of PF reliability indices is in certification and licensure examinations, which usually are taken by several thousand or at least several hundred examinees. Hence, investigating the bias of the procedures for large sample sizes seemed more important than comparing the small sample standard errors of the procedures.

The present simulations were undertaken on an IBM 4381 mainframe using SAS version 5 (SAS Institute Inc., 1985), except that the IMSL function BNRDF (International Mathematical and Statistical Libraries, 1987) was used to evaluate the bivariate normal cumulative distribution function for the normal method. Six simulation situations were considered, including three different test score distribution shapes (nearly normal, platykurtic, and negatively skewed) and two full-test reliabilities (.92 and .71). Full-test reliability was defined as the Spearman-Brown stepped-up correlation between the half-tests. For each of the six simulation situations, two replications were conducted in which one replication consisted of generating four half-test scores for 20,000 examinees under the following model:

$$Y_{ij} = \gamma^2 T_i + \gamma E_{ij} + \beta \tag{9}$$
$$(i = 1, \ldots, 20{,}000;\ j = 1, 2, 3, 4) \quad ,$$

where γ and β are parameters and T and all E_{ij} are independent variates generated from the standard normal distribution. All half-test scores were rounded to integer values, and the full-test scores were computed as $X_1 = Y_1 + Y_2$ and $X_2 = Y_3 + Y_4$. Various degrees of symmetrical and asymmetrical truncation on T, and to a much lesser extent symmetrical truncation on the Es, were used to control the distributional shapes of the simulated test scores. Formulas for the mean and variance of truncated normal variables are available (Johnson & Kotz, 1970). Manipulating the γ and β parameters in these formulas permitted some control over the means, variances, and reliabilities of the test scores.

Full-test characteristics for the six simulation situations are presented in Table 2. These situations were selected to represent those actually encountered in practice. Test score distributions for licensure, certification, and various other selection examinations are frequently—but not always—negatively skewed. Alternatively, Lord (1955) found that professionally constructed educational examinations resulted in score distributions that were symmetric and platykurtic (flat). Lord's (1955) finding was reaffirmed with a random sample of 40,000 examinees from a recent ACT Assessment examination (American College Testing Program, 1987). The distributions of raw scores for the four subtests comprising the ACT Assessment were approximately symmetric with skews ranging from −.2 to +.25 but with kurtoses ranging from −.40

Table 2
Reliability (ρ_{SB}) and Descriptive Statistics for the Simulation Distributions

Distribution	ρ_{SB}	Mean	SD	Skew	Kurtosis
Normal[a]	.92	160	23.0	.01	−.17
Normal[a]	.71	70	8.5	.00	−.14
Flat	.92	140	25.4	.01	−.93
Flat	.71	50	8.7	−.01	−.65
Skewed	.92	165	20.5	−.84	.47
Skewed	.71	57	5.5	−.63	.42

[a] These were nearly normal with slight platykurtosis.

to −.96. Finally, because test scores are inherently bounded, a nearly normal situation with slight platykurtosis was used instead of an exact (unbounded) normal distribution. Though full-test reliabilities for professionally constructed examinations are usually above .90 (or at least above .80), subtest reliabilities may be lower. Therefore, both high and low reliabilities represented by .92 and .71 were studied.

Failure rates of 10% and 30% were selected for investigation because they represented a realistic range. Due to the integer nature of the generated test scores, it was not always possible to achieve exactly 10% or 30% failure rates for the full tests. Rather, the failure rates ranged from 8.5% to 11% and from 29% to 32.5% across the six situations.

For an estimator T of some parameter ψ, bias is defined as $E(T) - \psi$. The parameters of interest are the PF reliability indices θ^* and $\phi^* = \kappa^*$. For the distributions modeled, the theoretical values of these parameters are not known. However, the simulations generated two full-test scores for all simulated examinees, and applying the method of moments (MOM) to the 2×2 table derived from the pairs of full-test scores yielded consistent estimates for the parameters. With $N = 20{,}000$ these consistent MOM estimates should, for practical purposes, accurately reflect the true parameter values. These estimates have no subscript and are denoted by a caret in the tables. The two estimation procedures were the normal and non-distributional methods, both of which were computed for the first full-test score $X_1 = Y_1 + Y_2$ only. In the tables, normal method estimates have N as a subscript while the non-distributional method estimates have SB as a subscript. The bias for each method is estimated as the difference between its estimate and the consistent MOM estimate. This approach of using the MOM estimate as the parameter is similar to that used by Huynh and Sanders (1980), Peng and Subkoviak (1980), and Subkoviak (1978).

Results

Table 3 presents MOM estimates for the PF reliability indices as well as estimated biases for the normal and non-distributional methods. Results are presented for two replications under all six simulation situations and for approximate fail rates of 10% and 30%.

The replications in Table 3 reveal some variability in the bias estimates, even with $N = 20{,}000$. Despite this variability, clear patterns emerge. Focusing first on θ^* shows that for the two nearly normal situations, the normal method is never significantly worse—and in one case it is appreciably better—than the non-distributional method, though the bias for both methods is modest. In the four non-normal situations, the pattern is reversed. The non-distributional method is never substantially worse and usually considerably better than the normal method, but again, both methods usually show only modest bias.

The pattern is similar for ϕ^*. However, the biases are generally larger, which Peng and Subkoviak (1980) and Huynh and Sanders (1980) also ob-

Table 3
Parameter Estimates and Bias Estimates for the Pass-Fail Reliability Indices by Distribution (Two Replications Each) With 10% Failing and 30% Failing for N = 20,000 and $\phi^* = \kappa^*$

Distribution	ρ_{SB}	Rep.	$\hat{\theta}^*$	$\hat{\theta}^*_{SB}-\hat{\theta}^*$	$\hat{\theta}^*_{N}-\hat{\theta}^*$	$\hat{\phi}^*$	$\hat{\phi}^*_{SB}-\hat{\phi}^*$	$\hat{\phi}^*_{N}-\hat{\phi}^*$
10% Failing								
Normal[a]	.92	1	.94	.007	−.003	.70	.034	−.009
		2	.95	.007	−.004	.71	.029	−.018
Normal[a]	.71	1	.90	.001	−.006	.41	.004	.006
		2	.90	.004	−.001	.40	.011	.016
Flat	.92	1	.92	.012	.028	.55	.086	.133
		2	.92	.009	.026	.57	.066	.114
Flat	.71	1	.88	.008	.018	.31	.053	.108
		2	.87	.012	.021	.29	.088	.124
Skewed	.92	1	.96	.009	−.007	.77	.041	−.089
		2	.96	.010	−.008	.77	.049	−.088
Skewed	.71	1	.93	.005	−.022	.56	.057	−.147
		2	.93	.000	−.026	.57	.030	−.156
30% Failing								
Normal[a]	.92	1	.89	.017	.000	.73	.040	.001
		2	.88	.021	.003	.72	.051	.008
Normal[a]	.71	1	.78	.008	−.005	.48	.026	.007
		2	.79	.013	−.009	.50	.035	−.006
Flat	.92	1	.90	.015	−.012	.77	.037	−.043
		2	.91	.016	−.014	.78	.044	−.050
Flat	.71	1	.79	.008	−.015	.53	.012	−.037
		2	.79	.003	−.014	.52	.000	−.035
Skewed	.92	1	.91	.016	−.030	.78	.038	−.047
		2	.91	.013	−.029	.78	.031	−.047
Skewed	.71	1	.80	.013	−.044	.55	.031	−.051
		2	.80	.012	−.037	.54	.027	−.038

[a] These were nearly normal with slight platykurtosis.

served with their methods. For the two nearly normal situations, the normal method is appreciably better than the non-distributional method, though the latter method performs reasonably well. With the four non-normal situations, the non-distributional method usually performs fairly well and is considerably better than the normal method, which has rather large bias when the fail rate is 10%.

Although the normal method yields both positive and negative bias estimates, the biases for the non-distributional method shown in Table 3 are always positive. This corresponds with the hypothesis noted above, which suggested that any bias would be positive and could be attributed to attenuation due to grouping. In addition, previous simulation studies by Huynh and Sanders (1980) and Peng and Subkoviak (1980) found that the HPS method and Huynh's beta-binomial method had biases similar in magnitude to those found for the present methods, though previous studies concentrated on short tests while this study focused on long tests.

Huynh's (1976) beta-binomial method was applied to the simulated data, but the results are not reported in detail because of their similarity to the present normal model estimates. For all six simulation situations and for both failure rates, the absolute differences between the beta-binomial estimates and the normal estimates never exceeded .02, and were usually less than .01 for both θ^* and $\kappa^* = \phi^*$. Furthermore, these small differences did not systematically produce less bias for the beta-binomial method. This was expected. Both Huynh

(1976) and Peng and Subkoviak (1980) presented theoretical and practical evidence suggesting that normal methods well approximate the beta-binomial method, especially when test length is long. (In applying the beta-binomial method, the number of items on each test was selected so that KR-21 was close in value to ρ_{SB}, with the constraint that test length could never be less than the maximum observed score.)

Thus, neither method showed large bias when estimating θ^*, though the normal method generally showed less bias than the non-distributional method when the test scores were approximately normally distributed; the opposite held when they were not. In estimating ϕ^*, the non-distributional method generally showed mild to moderate positive bias, but was considerably less biased than the normal method when the test scores were not normally distributed. When test scores were normally distributed, estimates of ϕ using the non-distributional method were moderately more biased than the normal method.

Example Application

The data used here were from a licensure examination containing 300 scored items. The test was divided into two separately timed parts consisting of 150 scored items each. The two parts were constructed to be equally difficult based on field test data, and were matched in content according to the test's table of specifications. A group of approximately 20 expert judges rated the 300 scored items using the Angoff (1984) method. The judges also specified what proportion of items a minimally competent examinee should answer correctly in each of the many content areas covered by the test. A passing score for the total test of at least 200 correct answers was determined from a weighted average of the judges' item and area ratings.

The method requires that passing scores be determined for the two parts. From a strictly statistical perspective, passing scores should be selected so that the passing rates on the two parts equal each other and the percentage passing on the full test. If a representative sample of examinees is available, then the half-test passing scores may be determined solely from the passing rates. The half-test passing scores need not be one-half of the full-test passing score, nor must the half-test passing scores sum to the full-test passing score. In general, these last two conditions will not be fulfilled when the half-test passing scores are determined by equating the passing rates.

In many applications, integrating psychometric and statistical considerations may be possible. If criterion data, such as expert judges' ratings, are available, then these data may also help to determine the half-test passing scores. Consider the present example: In rating the items and areas, the judges' only concern was establishing a passing score for the total test which would determine whether an examinee should be licensed. However, using the same weights as were used to determine the passing score for the total test, the item and area ratings were also used to determine passing scores for the two parts. Both parts received 100 items correct as passing scores based on the expert judges' ratings. After the test was administered and the results analyzed, these passing scores were changed to 102 for part 1 and 98 for part 2, because the average score for part 1 was approximately four points higher than for part 2. With these adjusted passing scores, the part 1 and part 2 passing rates were nearly identical to each other and to the full-test passing rate.

In this example, empirical results from a large representative sample were used to adjust the judges' ratings. Note that no decisions about examinees were based on the two half-test passing scores. Their only function was in estimating the full-test PF reliability. The fact that the half-test passing scores summed to the full-test passing score was due to the long length and corresponding high reliability of the two parts. If the two parts were shorter, this condition would likely have been violated.

Summary statistics for the total group and selected subgroups of examinees are presented in Table 4. The subgroup data illustrate the performance of the method for various sample sizes and different passing rates. Determination of the subgroups was based on whether an examinee was taking the test

Table 4
Summary Statistics for Each
Variable by Examinee Group

Group and Variable	Mean	SD	Skewness	Pass Rate
Total Group (N=4828)				
X	236.06	29.78	-1.41	.900
Y_1	119.93	15.11	-1.48	.894
Y_2	116.13	15.48	-1.24	.897
Accredited First Time (N=3999)				
X	241.96	22.59	-.80	.956
Y_1	122.97	11.39	-.88	.952
Y_2	118.99	12.20	-.69	.949
Accredited Repeating (N=548)				
X	224.01	28.73	-.27	.812
Y_1	113.59	14.79	-.39	.790
Y_2	110.42	14.84	-.17	.821
Non-Accredited First Time (N=94)				
X	187.22	49.61	-.14	.404
Y_1	95.01	24.28	-.19	.436
Y_2	92.21	26.06	-.17	.436
Non-Accredited Repeating (N=187)				
X	169.74	39.76	-.03	.209
Y_1	86.05	20.38	-.13	.198
Y_2	83.69	20.38	.01	.251

for the first time or was repeating the test, and whether an examinee graduated from an accredited or nonaccredited university.

The sample alpha coefficients (Table 5) were derived from scores on the examination's five subtests which differed in content, rather than from the item scores. This explains why the sample alphas are slightly smaller than the sample KR-21s. Because the subtests differed widely in length, average subtest scores (rather than total subtest scores) were used to compute the alphas.

Generally the stepped-up reliability coefficient, $r_{SB} = 2r(Y_1,Y_2)/[1 + r(Y_1,Y_2)]$, should be larger than KR-21, though that is not always the case (see Table 5), probably because of the long length of the test. In any case, the two reliability coefficients are quite similar in all subgroups.

The data in Table 4 show that although tests Y1 and Y2 have similar standard deviations, their means tend to differ; hence Y1 and Y2 are not precisely parallel tests. Moreover, the negative skewness coefficients suggest moderate to severe departures of the data from normal distributions.

However, the stepped-up phi coefficient, ϕ^*_{SB}, is based on the assumption that B_1, B_2, and A all have the same value of p. The pass rate column of Table 4 indicates how well the data satisfy this assumption. The passing scores for Y1 and Y2 were selected so that the assumption would be fulfilled in the total group of examinees. The assumption is met in the group of accredited first-time examinees but is violated to varying degrees in the remaining three groups.

As previously discussed, the observed proportions may be smoothed in the application of the non-distributional method. The "smoothed half-test proportions" in Table 6 were obtained by replacing the two off-diagonal proportions with their average. The estimated full-test proportions and PF reliability indices in Tables 6 and 7 were computed from the smoothed proportions.

The HPS estimates of the full-length PF reliability indices are based on Brennan's (1981) tables. Because KR-21 is nearly identical to the stepped-up reliability coefficient, r_{SB}, the HPS PF reliability indices are nearly identical to those that result from

Table 5
Reliability Coefficients by Examinee Group

Examinee Group	Alpha(X)	KR21(X)	r(Y_1,Y_2)	r_{SB}
Total group	.92	.95	.90	.94
Accredited				
First time	.87	.91	.84	.91
Repeating	.90	.93	.88	.94
Non-accredited				
First time	.96	.98	.94	.97
Repeating	.91	.96	.90	.95

applying the normal model to the half-test data. For this reason, the PF reliability indices associated with the normal model are omitted from Table 7.

Comparing the SB and HPS PF reliability estimates in Table 7 shows that they yield nearly identical estimates for θ^*, but that their estimates for κ^* are sometimes discrepant. The results from this example are consistent with the simulation results. When N is large and the test score distribution is substantially skewed, as in the first two groups in Table 7, the two methods give substantially different estimates for $\phi^* = \kappa^*$. The simulation results indicate that the SB method estimates should be more accurate, which is supported in this example by the similarity of the HPS estimates to the unstepped-up half-test MOM estimates of ϕ. Doubling the length of the test should increase ϕ at least a moderate amount; that was always the case in the simulations. The other three groups are considerably less skewed (with fairly normal kurtoses also) and in these groups the HPS and SB estimates for ϕ^* are more similar and substantially increased over the unstepped-up half-test MOM estimates of ϕ.

Conclusions

The methods for computing PF reliability presented here require only one test administration and use the Spearman-Brown formula to obtain stepped-up estimates of PF reliability which are computed from parallel half-tests. They thus require that the test be divisible into two parts equivalent in content and approximately equivalent in certain statistical characteristics. If this is not the case, then methods based on a beta-binomial model as discussed by Subkoviak (1980) could be used, such as the one by Huynh (1976). However, the beta-binomial method is computationally complex and seems more appropriate when tests are short and homogeneous

Table 6
Pass-Fail (PF) Proportions for Half-Test and Full-Test by Examinee Group

PF Decision		Half-Test Proportion		Full-Test
B_1	B_2	Raw	Smoothed	Proportion
Total Group				
0	0	.078	.078	.089
0	1	.028	.026	.015
1	0	.025	.026	.015
1	1	.870	.870	.881
Accredited First Time				
0	0	.030	.030	.037
0	1	.018	.019	.012
1	0	.021	.019	.012
1	1	.931	.931	.938
Accredited Repeating				
0	0	.133	.133	.156
0	1	.077	.061	.038
1	0	.046	.061	.038
1	1	.745	.745	.768
Non-Accredited First Time				
0	0	.500	.500	.527
0	1	.064	.064	.037
1	0	.064	.064	.037
1	1	.372	.372	.399
Non-Accredited Repeating				
0	0	.722	.722	.744
0	1	.080	.054	.032
1	0	.027	.054	.032
1	1	.171	.171	.193

Table 7
Reliability Indices by Examinee Group for Half-Test and Full-Test

Examinee Group	Half-Test $\hat{\phi}=\hat{\kappa}$	Half-Test $\hat{\theta}$	Full-Test $\hat{\phi}^*_{SB}=\hat{\kappa}^*_{SB}$	Full-Test $\hat{\theta}^*_{SB}$	Full-Test $\hat{\phi}^*_{HPS}=\hat{\kappa}^*_{HPS}$	Full-Test $\hat{\theta}^*_{HPS}$
Total group	.72	.95	.84	.97	.76	.95
Accredited						
First time	.59	.96	.74	.98	.61	.98
Repeating	.61	.88	.76	.92	.76	.92
Non-accredited						
First time	.74	.87	.85	.93	.90	.95
Repeating	.69	.89	.82	.94	.78	.92

in content and item difficulties. For long tests which are heterogeneous in content and item difficulties, such as licensure examinations, the Peng and Subkoviak (1980) approximation should yield results nearly identical to those from the beta-binomial method.

If the test is divisible into parallel half-tests, then the present methods possess certain advantages. Instead of KR-21, which is used in the HPS method, the normal method uses a Spearman-Brown stepped-up half-tests intercorrelation to estimate the correlation between two full tests. This full-test correlation estimate is based on less restrictive assumptions than those required by KR-21, and consequently the normal method may have wider applicability than the HPS method. In particular, it should be better suited to long heterogeneous tests such as licensure examinations, though this may not always be the case. Of more importance, however, is the non-distributional method which discards distributional assumptions altogether. The simulation results support the conclusion that when N is large and the test score distribution is non-normal, the non-distributional method will yield more accurate estimates than the normal method, especially for $\phi^* = \kappa^*$ and smaller (10%) failure rates.

Although the non-distributional method outperformed the normal method when normality was violated, it still displayed mild to moderate bias. The bias, however, was always positive, in contrast to the normal method, suggesting that it may be worthwhile to investigate strategies for correcting the bias. Also, the magnitudes of the biases found here were generally similar to those found by Peng and Subkoviak (1980) for their approximate method and to those found by Huynh and Sanders (1980) for the beta-binomial method, though these two studies focused on short tests. Finally, the simulation results obtained here only apply to large sample sizes. Investigating the behavior of the non-distributional method when sample size is small, test length is short, and test score distributions are non-normal could extend the applicability of the methods to situations other than the professional licensure and certification examinations considered here.

References

American College Testing Program. (1987). *ACT assessment technical manual*. Iowa City IA: Author.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington DC: Author.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton NJ: Educational Testing Service. [Reprint of chapter in R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington DC: American Council on Education, 1971.]

Brennan, R. L. (1981). *Some statistical procedures for domain-referenced testing: A handbook for practitioners* (ACT Technical Bulletin No. 38). Iowa City IA: American College Testing Program.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

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

Huynh, H. (1976). On the reliability of decisions in domain referenced testing. *Journal of Educational Measurement, 13,* 253–264.

Huynh, H., & Sanders, J. C. (1980). Accuracy of two procedures for estimating reliability of mastery tests. *Journal of Educational Measurement, 17,* 351–358.

International Mathematical and Statistical Libraries. (1987). *STAT/LIBRARY: FORTRAN subroutines for statistical analysis* (Vol. 3). Houston: Author.

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions—1.* Boston: Houghton Mifflin.

Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement, 15,* 383–389.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359–368.

SAS Institute Inc. (1985). *SAS user's guide basics, version 5 edition.* Cary NC: Author.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement, 15,* 111–116.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 267–291). Baltimore MD: Johns Hopkins University Press.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25,* 47–55.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision theoretic formulation. *Journal of Educational Measurement, 11,* 263–267.

Author's Address

Send requests for reprints or further information to David J. Woodruff or Richard L. Sawyer, American College Testing Program, P.O. Box 168, Iowa City IA 52243, U.S.A.