DOCUMENT RESUME

ED 382 655 TM 023 089

| | |
|---|---|
| AUTHOR | Breyer, F. Jay; Lewis, Charles |
| TITLE | Pass-Fail Reliability for Tests with Cut Scores: A Simplified Method. |
| INSTITUTION | Educational Testing Service, Princeton, N.J. |
| REPORT NO | ETS-RR-94-39 |
| PUB DATE | Jul 94 |
| NOTE | 39p. |
| PUB TYPE | Reports - Evaluative/Feasibility (142) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *Classification; *Cutting Scores; Objective Tests; *Pass Fail Grading; *Reliability; *Research Methodology; Statistical Analysis; Test Construction |
| IDENTIFIERS | Split Half Test Reliability |

ABSTRACT

A single-administration classification reliability index is described that estimates the probability of consistently classifying examinees to mastery or nonmastery states as if those examinees had been tested with two alternate forms. The procedure is applicable to any test used for classification purposes, subdividing that test into two half-tests, each with a cut score, where the sum of the two half-test cut scores is equal to the cut score for the total test. The application of this pass-fail consistency index to binary scored objective tests, nonbinary scored performance tests. and tests containing both binary and nonbinary scored questions is presented. A calculation example is provided together with look-up tables. (Contains 13 references, 9 tables, and 3 figures.) (Author)

RR-94-39



RESEARCH REPORT



# PASS-FAIL RELIABILITY FOR TESTS WITH CUT SCORES: A SIMPLIFIED METHOD

F. Jay Breyer
Charles Lewis

Educational Testing Service
Princeton, New Jersey
July 1994

# Pass-Fail Reliability for Tests with Cut Scores:

## A Simplified Method

F. Jay Breyer and Charles Lewis

*Educational Testing Service*

R JNNING HEAD: PASS-FAIL RELIABILITY FOR TESTS WITH CUT SCORES

## Abstract

A single-administration classification reliability index is described that estimates the probability of consistently classifying examinees to mastery or nonmastery states as if those examinees had been tested with two alternate forms. The procedure is applicable to any test used for classification purposes, subdividing that test into two half-tests, each with a cut score, where the sum of the two half-test cut scores is equal to the cut score for the total test. The application of this pass-fail consistency index to binary scored objective tests, nonbinary scored performance tests, and tests containing both binary and nonbinary scored questions is presented. A calculation example is provided together with look-up tables.

Pass-Fail Reliability for Tests with Cut Scores: A Simplified Method

It is common practice for industrial psychologists to use tests that have cut scores. The essential feature of these tests is that the interpretive meaning is based upon an examinee's obtained score with regard to some cut or passing score. Tests with cut scores take many forms. Some contain mostly objective items (e.g., multiple choice items), other tests are composed entirely of scored performances, and still others combine objective items with performance-based tasks that measure an examinee's knowledge, skill and/or proficiency in real-life situations. Examples of scored performances include in-basket tests (Frederiksen, 1960; Frederiksen, 1962; Frederiksen Jensen, & Beaton, 1972; Frederiksen, Saunders, & Wand, 1957), constructed response or free response tests (Bennett, 1993), figural response tests (e.g., drawings), and essay tests. Many of these tests have some questions or items scored dichotomously (e.g., 0,1) and others that receive a range of scores (e.g., -0.5, 0, 2.3 or 0, 1, 2, 3, 4, 5).

Even though composite tests with a mix of item types are being used to make individual decisions, and thus have cut scores for classifying examinees into mastery or nonmastery states, coefficient alpha is still the most commonly reported reliability estimate (Cortina, 1993). If a test is used to classify examinees into mastery or nonmastery states, then the appropriate error of measurement has to do with inconsistently classifying individuals. Alpha is best for assessing errors of measurement related to content heterogeneity, not for evaluating alternate form classification consistency. In other words, the question of interest should be: Would the same people pass and fail if they were tested with an alternate form? The misuse of alpha as the sole consistency estimate for tests with cut scores may be due to the lack of accessibility of an appropriate

classification consistency index for tests with a combination of differently scored items as opposed to the readily available alpha and its derivatives.

## What is Available to Assess Pass-Fail Consistency?

Tests with a combination of differently scored items can present problems for professionals in calculating reliability statistics related to placing individuals into mastery or nonmastery (pass-fail) categories based on test scores that are sums of dichotomous and polytomous item scores. One pass-fail reliability estimate, a single-administration classification consistency index (Subkoviak, 1976, 1988), estimates the consistency of passing or failing the same individuals if each were given the same test form twice. This classification consistency index is thought to work with dichotomously scored questions only (e.g., rights scored multiple choice items and short answer questions) because of the assumption of either a binomial model or a compound binomial model underlying the score distributions of alternate forms (Subkoviak, 1976). However, if one makes a strong assumption of bivariate normality between score distributions of alternate forms, Subkoviak's index can be used to assess the alternate form classification consistency for tests with nonbinary scored questions.

A second, less well-known method is the Relclass procedure for estimating classification consistency (Livingston & Lewis, in press). A frequency distribution of observed scores, an estimate of the reliability of the test (such as alpha), and a cut score for classification purposes are needed. In Relclass, a true score distribution is fitted to the observed score data. The way this is accomplished is to assume that the error distribution is binomial. Employing these two distributions, a bivariate distribution for the observed scores and those from a (hypothetical) replication of the test is constructed. Using the cut score on the observed scores and the replication, this bivariate distribution is divided into four

parts, and the probability associated with each of the quadrants is evaluated. Summing the estimated probabilities for pass-pass and fail-fail gives an estimate of the probability of consistent classification.

A third option is a new statistical procedure for estimating the probability of consistently classifying examinees to mastery or nonmastery states for test score data from one administration. This pass-fail consistency index can be computed with test scores composed of binary scored items, nonbinary scored items, or any combination where the test scores are used for classification (i.e., pass-fail) purposes. The new reliability index procedure requires test score data to be subdivided into scores on two half-tests (which should be comparable, but do not have to be strictly parallel), each with a cut score, where the sum of the two half-test cut scores is equal to the cut score for the total test. A typical choice for the half-test cut scores would be to divide the full-test cut score by two.

First, a description of this new pass-fail consistency index is provided. Second, empirical results in applying the estimate are presented and compared to other (appropriate and inappropriate) reliability estimates for assessing pass-fail consistency. Third, a step-by-step numerical illustration of the calculation of such an index is provided, together with look-up tables for practical use.

## A Split-Half Pass-Fail Consistency Index

Subkoviak (1976) provides a simple formula (his Equation 11, our Equation 1) to estimate the probability of consistent classifications to be used when the observed score distribution is approximately normal and a reliability estimate (such as alpha) is available. It is

$$P_{cc} = 1 - 2[P(z < c) - P(z < c, z' < c)] \quad , \tag{1}$$

where $P(z < c)$ is the proportion of observed scores below the cut score, and

$P(z < c, z' < c)$ is the probability from a bivariate normal distribution with correlation equal to the reliability, that both variables are less than the (standardized) cut score.

Conside. now a situation where either the observed score distribution is not close to normal (e.g., licensure and/or certification test score distributions that are often negatively skewed) or no internal consistency reliability estimate is available. Suppose instead that the full test can be split into two comparable half-tests, each with its own cut score, so that a 2x2 pass-fail table of frequencies can be constructed for the half-tests. It is assumed that there is an underlying bivariate normal distribution associated with this 2x2 table (see Figure 1). This is <u>not</u> the same as assuming that the two half-tests have a bivariate normal distribution. Instead, it is essentially the assumption on which the computation of a tetrachoric correlation is based. Another way of stating the assumption is that doubling the test length will affect the 2x2 pass-fail tables based on the actual scores in the same way that it affects 2x2 tables based on the bivariate normal distribution.

The basic idea of the method described is to estimate the tetrachoric correlation for the half-test 2 x 2 table, take it as an estimate of the 'effective' reliability of the half-tests, apply the Spearman-Brown prophecy formula to estimate the effective reliability of the full test and apply Subkoviak's Equation 11 to obtain an estimate of the probability of consistent classifications for the full test.

==========

Insert Figure 1 about here

==========

The first step in computing the new index is to split the test into comparable half-tests, each with a half-test cut score, and construct a 2 x 2 contingency table as shown in Figure 1.

Second, calculate the average proportion of failures for the two half-tests as

$$P_{f,half} = \frac{2X_{11} + X_{12} + X_{21}}{2N}, \tag{2}$$

where $X_{11}$ is the frequency of examinees failing half-test 1 and half-test 2; $X_{12}$ and $X_{21}$ are the frequencies of examinees failing half-test 1 but passing half-test 2, and passing half-test 1 but failing half-test 2, respectively; and $N$ is the total number of examinees taking the test in one administration. (See Figure 1.)

Third, find the standard normal deviate ($z_{half}$) for which the cumulative normal probability is equal to $P_{f,half}$. (See Figure 2.) Consult any introductory statistics text with a normal curve table.

==========

Insert Figure 2 about here

==========

The proportion of examinees who fail both half-tests ($P_{ff,half}$) is

$$P_{ff,half} = \frac{X_{11}}{N} \quad . \tag{3}$$

Using the $z_{half}$ and the $P_{ff,half}$ values, the correlation between the two hypothetical half-tests [ those with the assumed bivariate normal distribution] is estimated. (This is actually a special case of estimating a tetrachoric correlation.) The correlation can be found by using (in reverse) the tables summarized in Huynh

(1976) and fully described in Gupta (1963). It may also found as the value of $r_{half}$, which makes the following equation true:

$$P_{ff,half} = \int_{-\infty}^{z_{half}} \int_{-\infty}^{z_{half}} \frac{1}{2\pi\sqrt{1-r_{half}^2}} \exp\left[-\frac{1}{2}\frac{\left(z_1^2 - 2r_{half}z_1z_2 + z_2^2\right)}{\left(1-r_{half}^2\right)}\right] dz_1 dz_2 \quad , \tag{4}$$

where $z_1$ and $z_2$ are variables of integration corresponding to each half test.

This special case of the tetrachoric correlation ($r_{half}$) is then stepped up with the Spearman-Brown prophesy formula to find the reliability for the hypothetical full test,

$$r_{full} = \frac{2r_{half}}{(1+r_{half})}. \tag{5}$$

The standard deviation of the sum of the two hypothetical half-tests is then estimated as

$$s_{full} = \sqrt{1+1+2r_{half}}. \tag{6}$$

The sum of the cut scores for the hypothetical half-tests is restandardized for the full test as shown in Equation 7,

$$z_{full} = \frac{2z_{half}}{s_{full}}, \tag{7}$$

and the probability of failing the full test ($P_{f,full}$) is estimated by taking the cumulative standard normal probability corresponding to $z_{full}$.

The $z_{full}$ and the $r_{full}$ are employed to estimate the probability of failing two full tests ($p_{ff,full}$) through use of Huynh's (1976) tables or integration corresponding to Equation 4 with the replacement of $r_{half}$ with $r_{full}$. Finally,

Subkoviak's (1976) Equation 11, or our Equation 1, is used to estimate the probability of consistent classification for the full test:

$$P_{cc,full} = 1 - 2(P_{f,full} - P_{ff,full}) \tag{8}$$

## How Well Does This Index Work in Practice?

The pass-fail consistency index was evaluated for tests with binary scored questions, non-binary scored questions and a mix of both binary and non-binary scored questions and compared against existing pass-fail consistency indexes and other internal consistency indexes. First, the pass-fail consistency index was compared with the Subkoviak index and the Relclass consistency index for three dichotomously scored multiple choice tests (study 1). Two polytomously scored in-basket tests (study 2) were also compared with existing pass-fail consistency indexes and alpha, and two mixed scored composite assessments consisting of multiple-choice and constructed response questions (study 3) were examined. To determine how sensitive the index is to the assumption of parallelism of the half-tests, two in-basket half-tests were further subdivided into (non-parallel) quarter-tests (study 4) and the resulting half-test reliability indexes were compared to the empirical results of the proportion of consistent classifications for the half tests.

## Study 1

### Method

Three four-choice 75-item knowledge tests used for licensure decisions were evaluated with the new pass-fail consistency index. Each of the items on these tests was scored either "1" for right responses or "0" for wrong responses. All tests were evaluated with coefficient alpha, the Subkoviak pass-fail consistency index, Relclass, and the new pass-fail classification consistency index requiring half-tests each with their own cut score. Each of these multiple choice tests was

divided into odd-even half-tests and assigned a cut score equal to the full-test cut score divided by two, with the exception of the third test. The third test had an odd total test cut score that prohibited the construction of half-tests each with the identical cut score, thus the odd-even half-tests were evaluated in two different ways -- with cut scores that summed to the total test cut score but were not identical -- and the results of these two consistency index computations were averaged. Note that the 75-item tests do not divide into half-tests with equal numbers of items.

Results

Table 1 gives the results of the reliability comparisons for the three dichotomously scored multiple choice tests.

==========

Insert Table 1 about here

==========

Study 2

Method

Two nonbinary scored constructed response tests, in-basket tests (Frederiksen, 1960; Frederiksen, 1962; Frederiksen Jensen, & Beaton, 1972; Frederiksen, Saunders, & Wand, 1957), each with a cut score used for voluntary certification decisions were evaluated. Each in-basket test consists of a number of scorable units, i.e., problems or issues, that are embedded in written documents such as memos, letters, telephone and e-mail messages and reports. (Test #4 consisted of 60 problems embedded in 22 documents and test #5 consisted of 49 problems in 20 documents.) Each problem was scored on a continuum that ranged from -0.5 through +1.0; total decision scores were computed by summing the problem

scores. Problems contained in one document, and problems that related to problems in other documents were grouped into the same half test to avoid inflating the half-test correlation and the estimated classification consistency. See Sireci, Wainer, and Thissen (1991) for a demonstration of how common stimuli, such as a reading passage, can inflate reliability coefficients based on item scores.

Results

Table 2 gives the results of the reliability comparisons for the two non-binary scored in-basket tests.

==========

Insert Table 2 about here

==========

Study 3

Method

Two composite assessments were evaluated, each composed of a binary scored multiple choice test and a non-binary scored constructed response test -- essays -- with one total cut score used for advanced placement decisions in college course sequences. These composite assessment scores are reported on a scale ranging from a low score of 1 through a high score of 5. The measure of internal consistency evaluated in such a composite assessment is the reliability of the composite raw scores, given as

$$r_{composite} = 1 - \frac{\sum_{i=1}^{k} w_i^2 SE_i^2}{\sigma_X^2}, \qquad (9)$$

where $k$ is the number of test parts contributing to the composite score, $w_i$ is the weight applied to test part $i$, $SE_i^2$ is the square of the standard error of

measurement for each test part $i$, and $\sigma_X^2$ is the variance of the composite scores. See Feldt and Brennan (1989) for further information.

Because different schools have different cut score limits for granting advanced placement standing, the pass-fail consistency reliability is calculated for four different cut points on the scale score range. (Few colleges grant advanced placement below a scale score of three.) Composite half tests were constructed by placing essay scores that belong to one question in the same half test so that examinee responses to essay questions with multiple scores did not artificially inflate the alternate form reliability.

Results

Table 3 gives the results of the reliability comparisons for the two composite tests. Note how the classification consistency indexes vary depending on the pass rate, actually where the cut score is in the distribution, with those classification consistency indexes at cuts between scale scores of 4 and 5 and between scale scores of 1 and 2 showing higher classification consistencies than those cut scores in the middle of the distribution.

==========

Insert Table 3 about here

==========

Study 4

Method

One half-test from each of the in-basket examinations (test #4 and test #5) was subdivided into two tests, each containing a quarter of the scorable units, by separating those problems dependent on the same or related documents into the same quarter-length test and those problems dependent on other documents into the other quarter-length test. This procedure of further subdividing a half-test

into two quarter-length tests assured that different content was measured in each quarter-length test. These quarter-length tests were then used to determine the classification consistency for alternate forms of the half-tests. Thus, the consistent classification probability estimates of the new index could be compared to the actual proportion of consistent classifications obtained for half-tests. The proportion of consistent classifications ($P_{cc\ half}$) is calculated as

$$P_{cc\ half} = \frac{X_{22} + X_{11}}{N}, \tag{10}$$

where $X_{22}$ indicates the frequency of passing candidates on half-test 1 and half-test 2, $X_{11}$ is the frequency of examinees failing half-test 1 and half-test 2 and $N$ is the total number of examinees who took the test in one administration.

Results

Table 4 shows the actual proportions of consistent classifications from the in-basket half-tests and the estimated probabilities of consistent classifications based on the new index from quarter-length tests. Note the similarity of the estimates to the observed proportions.

==========

Insert Table 4 about here

==========

Discussion

The three procedures used to estimate classification consistency, (1) the Subkoviak, (2) Relclass, and (3) the new classification consistency index, generally give similar results. These classification consistency indexes might begin to differ from one another when the score distributions for the alternate forms are seriously different from what is assumed or when other assumptions are badly violated. A review of the assumptions of each is in order.

An advantage of the Subkoviak index when using the normality assumption (compared with other approaches proposed by Subkoviak) is that it does not require binary test items. Binary scored test items are only required when using a binomial or compound binomial model in the computation of the index. A disadvantage of the Subkoviak index is that it makes a strong assumption of bivariate normality between the score distributions of the alternate forms and in some cases (i.e., licensure tests, and selection tests), where examinees have been preselected, score distributions are anything but normal. A second disadvantage is that it requires some form of a reliability estimate; typically alpha is used. See Cortina (1992) for a description of problems associated with alpha.

The major advantage of the Relclass index is that it works with test scores regardless of the composition of those scores. Relclass is useful for tests composed of binary scored items, and nonbinary scored items, as well as composite scores of independently scored tests. Disadvantages of Relclass are that it is not widely available, it is difficult to compute, and it requires a reliability estimate such as alpha. Also, Relclass assumes a unimodal true score distribution (Livingston & Lewis, in press), which might be problematic for some tests.

Advantages of the new classification consistency index, $P_{cc,full}$, are that it is relatively simple to compute, it makes a weaker assumption than the Subkoviak procedure about the distribution of scores from alternate test forms, and it makes direct use of a classification consistency table. The assumption of normality of the test score distribution is not made, only the assumption of normality of an underlying distribution. Again, this is much like the assumption made in the definition of the tetrachoric correlation. A disadvantage is that it requires that the

full test be subdivided into half-tests with comparable content and requires half-test cut scores. The construction of half-tests from tests composed of individually scored questions based on common stimulus materials (e.g., an architectural design problem or a common reading passage) requires those questions be assigned to the same half-test to avoid spurious inflation of the index. The classification consistency index, $P_{cc,full}$, is useful when the normality of test scores is not certain (or when test scores are certainly not normal) and when the reliability estimate is to be doubted (e.g., Kuder-Richardson formula 21 is all that is available for a test measuring heterogeneous content).

In any event, pass-fail consistency across alternate forms is not reflected in alpha or in other internal consistency estimates (i.e., the reliability of the composite). Whichever of the three indexes for the reliability of classification is used, do not use alpha. Alpha answers a different question that does not directly address passing or failing the same examinees on an alternate test form.

## A Calculation Example

Step 1. Sample data are presented in Figure 3 for illustrative purposes, where the pass-fail consistency procedure is applied to hypothetical data from a test used to assign candidates to either pass or fail classifications. As can be seen, a fourfold table is created from half-test data with each half-test possessing a cut score. Table 5 outlines the step-by-step calculation using Equations 2 through 8. However, for practical use, look-up tables are also provided. These look-up tables (Tables 7, 8, and 9) give the probabilities of interest in terms of the log odds for passing and the log odds ratio of consistency to inconsistency. This choice allows adequate coverage of the cases of interest with equal spacing of the rows and columns, thus facilitating interpolation in the tables.

Using the frequencies from Figure 3 (or Table 6a) and Equation 2, the average proportion of failures can be computed as $P_{f,half} = \frac{2X_{11} + X_{12} + X_{21}}{2N}$ or $0.268 = \frac{2 \times 13 + 12 + 14}{2 \times 97}$.

==========

Insert Figure 3 and Table 5 about here

==========

Step 2. For ease of computation, the interested reader is invited to change the frequencies (see Table 6a) to proportions, where the sum of the four cells equals 1.00 as shown in Table 6b.

Step 3. Using the average proportion of failures, $P_{f,half} = 0.268$, as the marginal table entry (see Table 6c), the proportions from the inconsistent cells within the table are adjusted so the average proportion of failures is evenly divided between them without modifying the consistent (fail, fail) cell.

==========

Insert Table 6 about here

==========

Step 4. Using the marginal proportions from Table 6c, the log odds for passing ($\ln(\omega_p)$) are calculated by taking the log of the ratio of the average proportion passing to the average proportion of failures: $\ln(\omega_p) = \ln\left(\frac{P_{p,half}}{P_{f,half}}\right) = \ln\left(\frac{0.732}{0.268}\right) = 1.00$. (Note that in cases where the proportion passing is less than the proportion failing, as in highly selective situations, simply reverse the odds of $\frac{P_{p,half}}{P_{f,half}}$ to $\frac{P_{f,half}}{P_{p,half}}$ to calculate the log odds for failures and insert this value where $\ln(\omega_p)$ is employed.) Next, using the cell entries in Table

6c, the log odds ratio of consistent to inconsistent classifications ($\ln(\omega_{c/i})$) is calculated as $\ln(\omega_{c/i}) = \ln\left(\frac{P_{pp,half} \times P_{ff,half}}{P^2_{fp,half}}\right) = \ln\left(\frac{0.134 \times 0.598}{0.134^2}\right) = 1.50$.

Step 5. With the two log odds obtained from step 4, the probability for passing the full test twice can be determined from Table 7 ($P_{pp,full}$ =.660) and placed in Table 6d as the cell entry for the (pass, pass) condition. In the same way, the marginal pass rate for two alternate forms may be found ($P_{p,full}$ =.761) in Table 8 and the remaining cells filled in as shown in Table 6d. The sum of the consistent cell entries should add to the total pass-fail classification consistency index $P_{cc,full} = P_{ff,full} + P_{pp,full} = 0.138 + 0.660 = 0.798$. This agrees with the final result given in Table 5. (Some of the intermediate results in Tables 5 and 6 differ slightly, due to rounding.)

Step 6. For those individuals not requiring the alternate forms estimation of the four cell entries, step 5 may be skipped. Table 9 provides direct evaluation of the pass-fail consistency estimate. The interested reader can verify that the intersection of the log odds of passing ($\ln(\omega_p)$=1.00) and the log odds ratio of consistent to inconsistent classifications ($\ln(\omega_{c/i})$=1.50) yields a pass-fail consistency estimate for the full test ($P_{cc,full}$) of 0.798, the same value obtained in step 5 after adding the two consistent cell proportions.

In step 4, the formulas used to compute the log odds of passing and the log odds ratio of consistent to inconsistent classifications are given. To construct Tables 7, 8 and 9, it was necessary to solve these for $P_{f,half}$ and $P_{ff,half}$, so that the calculation steps summarized in Table 5 could be followed. For reference, the formulas used were

$$P_{f,half} = \frac{1}{1+\omega_p} \tag{11}$$

and

$$P_{ff,half} = P_{f,half} + \frac{1-\sqrt{1+4(\omega_{c/i}-1)P_{f,half}(1-P_{f,half})}}{2(\omega_{c/i}-1)}, \qquad (12)$$

assuming $\omega_{c/i} > 1$.

Conclusion

It is hoped that this new pass-fail classification consistency index, $P_{cc,full}$, will help professionals who use tests with cut scores determine the consistency of the pass-fail decisions made with those tests. In such cases where tests are not homogeneous or internally consistent due to the kinds of knowledge, skill or proficiency required by complex jobs, this index should be helpful in answering the question: How reliable is this test for its intended use of classification? Score reliability estimates, notably coefficient alpha and its derivatives, are not appropriate and, in fact, might lead to the wrong conclusions for tests where the question concerns consistency of classification.

References

Bennett, R. E. (1993). Constructed-response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests*. (pp. 99-124). Hillsdale, NJ: Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An explanation of theory and applications. *Journal of Applied Psychology*, **78**, 98-104.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 105-146). New York: American Council on Education and Macmillan Publishing Company.

Frederiksen, N. (1960). In-basket tests and factors in administrative performance. In The Proceedings of the Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 208-221.

Frederiksen, N. (1962). Factors in in-basket performance. *Psychological Monographs: General and Applied*, **76**, (22, whole no. 541).

Frederiksen, N., Jensen, O., & Beaton, A. E. (1972). *Prediction of Organizational Behavior*. New York: Pergamon.

Frederiksen, N., Saunders, D. R., & Wand, B. (1957). The in-basket test. *Psychological Monographs: General and Applied*, **71**, (9, whole no. 438).

Gupta, S. S. (1963). Probability integrals of multivariate normal and multivariate t. *Annals of Mathematical Statistics*, **34**, 792-828.

Huynh, H. (1976). On consistency of decisions in criterion-referenced testing. *Journal of Educational Measurement*, **13**, 253-264.

Livingston, S. A., & Lewis, C. (in press). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, **28**, 237-247.

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

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement,* **25**, 47-56.

Table 1

Comparison of Reliability Indexes for Three Binary Scored Tests with Cut Scores

| Test | Score Range | Cut Score | N | Pass Rate | Alpha | Subkoviak | Relclass | $P_{cc,full}$ |
|---|---|---|---|---|---|---|---|---|
| #1 | 0 - 75 | 46 | 472 | 70% | .88 | .86 | .87 | .84 |
| #2 | 0 - 75 | 34 | 438 | 79% | .80 | .86 | .83 | .83 |
| #3 | 0 - 75 | 37 | 474 | 69% | .77 | .81 | .80 | .83[a] |

[a]Average of .809, based on cut of 19 for even items and 18 for odd items, and .852, based on cut of 18 for even items and 19 for odd items.

Table 2

Comparison of Reliability Indexes for Two Non-Binary Scored Tests with Cut Scores

| Test | Score Range | Cut Score | N | Pass Rate | Alpha | Subkoviak | Relclass | $P_{cc,full}$ |
|---|---|---|---|---|---|---|---|---|
| #4 | -30 - 60 | 28 | 181 | 64% | .64 | .74 | .77 | .73 |
| #5 | -24 - 49 | 22 | 108 | 71% | .63 | .74 | .80 | .75 |

Table 3

Comparison of Reliability Indexes for Two Composite Tests with one Total Cut

| Test | Score Range | Cut Score | N | Pass Rate | Composite Reliability | Subkoviak | Relclass | $P_{cc,full}$ |
|---|---|---|---|---|---|---|---|---|
| #6 | 1 - 5 | 4/5 | 2,485 | 11% | .93 | .95 | .96 | .95 |
| | | 3/4 | | 30% | | .89 | .92 | .90 |
| | | 2/3 | | 64% | | .89 | .91 | .89 |
| | | 1/2 | | 80% | | .92 | .92 | .92 |
| #7 | 1 - 5 | 4/5 | 113,129 | 11% | .80 | .90 | .90[a] | .90 |
| | | 3/4 | | 30% | | .82 | .82[a] | .80 |
| | | 2/3 | | 68% | | .82 | .82[a] | .81 |
| | | 1/2 | | 97% | | .97 | .97[a] | .98 |

[a]Relclass results are based on a subsample of 9,995 examinees for test #7.

Table 4

Comparison of Reliability Indexes

for Non-Binary Scored Half-Tests and Nonparallel Quarter-Tests with Cut Scores

| Test | Cut | Test Split | Estimated Probability of Consistent Classification for Half-Tests |
|---|---|---|---|
| #4 | 14 | Half-test (observed $P_{cc,half}$) | .67 |
| #4 | 7 | Quarter Test (estimate of $P_{cc,full}$) | .71 |
| #5 | 11 | Half-test (observed $P_{cc,half}$) | .69 |
| #5 | 5.5 | Quarter Test (estimate of $P_{cc,full}$) | .67 |

Table 5

Step by Step Calculation of the Pass-Fail Consistency Index ($P_{cc,full}$)

| Step | Procedural Calculation | Example |
|---|---|---|
| 1 | Split the test into two half-tests each with a half-test cut. | See Figure 3 |
| 2 | Calculate the average proportion of failures (Equation 2). | $P_{f,half} = .268$ |
| 3 | Find the z score corresponding to $P_{f,half}$. | $z_{half} = -0.619$ |
| 4 | Find the proportion of examinees who fail both half-tests (Equation 3). | $P_{ff,half} = .134$ |
| 5 | Using integration with $z_{half}$ and $P_{ff,half}$, find $r_{half}$ (Equation 4). | $r_{half} = .506$ |
| 6 | Using $r_{half}$ and the Spearman-Brown formula (Equation 5) find $r_{full}$. | $r_{full} = .672$ |
| 7 | Calculate the standard deviation for the full test, $s_{full}$, with Equation 6. | $s_{full} = 1.736$ |
| 8 | Use Equation 7 to calculate $z_{full}$ using the z score from step 3 and the $s_{full}$ from step 7. | $z_{full} = -0.713$ |
| 9 | Using a normal curve table, find $P_{f,full}$ corresponding to $z_{full}$. | $P_{f,full} = .238$ |
| 10 | Find $P_{ff,full}$ using integration with $z_{full}$ and $r_{full}$. | $P_{ff,full} = .137$ |
| 11 | Insert $P_{f,full}$ and $P_{ff,full}$ into Equation 8 to calculate the index, $P_{cc,full}$. | $P_{cc,full} = .798$ |

Table 6

Calculation of P/F Consistency with Sample Test Data

(a) Half-test Frequencies

| | Fail | Pass | |
|---|---|---|---|
| Fail | 13 | 12 | |
| Pass | 14 | 58 | |
| | | | 97 |

(b) Half-test Proportions

| | Fail | Pass | |
|---|---|---|---|
| Fail | 0.134 | 0.124 | |
| Pass | 0.144 | 0.598 | |
| | | | 1.00 |

(c) Adjusted Proportions

| | Fail | Pass | |
|---|---|---|---|
| Fail | 0.134 | 0.134 | |
| Pass | 0.134 | 0.598 | |
| | 0.268 | 0.732 | 1.00 |

(d) Full-test Proportions

| | Fail | Pass | |
|---|---|---|---|
| Fail | 0.138 | 0.101 | |
| Pass | 0.101 | 0.660 | |
| | 0.239 | 0.761 | 1.00 |

Table 7

Estimated Probability of Passing Two Alternate Forms ( $P_{pp,full}$ )

| $\ln[\omega_p]$ | Log Odds Ratio Consistent:Inconsistent, $\ln[\omega_{c/i}]$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Log Odds Passing | .25 | .50 | .75 | 1.00 | 1.25 | 1.50 | 1.75 | 2.00 | 2.25 | 2.50 | 2.75 | 3.00 |
| .00 | .279 | .303 | .323 | .342 | .358 | .373 | .385 | .397 | .408 | .417 | .425 | .433 |
| .25 | .368 | .387 | .404 | .419 | .433 | .445 | .456 | .467 | .476 | .484 | .492 | .499 |
| .50 | .464 | .477 | .489 | .500 | .510 | .519 | .528 | .536 | .544 | .551 | .557 | .563 |
| .75 | .560 | .566 | .572 | .578 | .585 | .591 | .598 | .604 | .610 | .615 | .621 | .626 |
| 1.00 | .651 | .650 | .651 | .653 | .656 | .660 | .663 | .667 | .671 | .676 | .680 | .684 |
| 1.25 | .732 | .727 | .723 | .722 | .722 | .722 | .724 | .726 | .728 | .730 | .733 | .736 |
| 1.50 | .800 | .792 | .786 | .782 | .779 | .778 | .777 | .777 | .778 | .779 | .781 | .783 |
| 1.75 | .855 | .846 | .839 | .833 | .829 | .826 | .824 | .822 | .822 | .822 | .823 | .824 |
| 2.00 | .898 | .889 | .882 | .875 | .870 | .866 | .863 | .861 | .859 | .858 | .858 | .858 |
| 2.25 | .929 | .922 | .915 | .909 | .904 | .899 | .895 | .892 | .890 | .889 | .888 | .887 |
| 2.50 | .952 | .946 | .940 | .935 | .930 | .925 | .922 | .918 | .916 | .914 | .912 | .912 |
| 2.75 | .968 | .963 | .959 | .954 | .950 | .946 | .942 | .939 | .936 | .934 | .933 | .931 |
| 3.00 | .979 | .976 | .972 | .968 | .965 | .961 | .958 | .955 | .953 | .951 | .949 | .947 |

Table 8

Estimated Probability of Passing the Full Test ($P_{p,full}$)

| $\ln[\omega_p]$ | Log Odds Ratio Consistent:Inconsistent, $\ln[\omega_{c/i}]$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Log Odds Passing | .25 | .50 | .75 | 1.00 | 1.25 | 1.50 | 1.75 | 2.00 | 2.25 | 2.50 | 2.75 | 3.00 |
| .00 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 | .500 |
| .25 | .584 | .580 | .577 | .575 | .573 | .571 | .569 | .568 | .567 | .566 | .565 | .565 |
| .50 | .663 | .657 | .652 | .647 | .643 | .639 | .637 | .634 | .632 | .630 | .629 | .628 |
| .75 | .735 | .727 | .720 | .714 | .708 | .703 | .699 | .696 | .693 | .690 | .688 | .687 |
| 1.00 | .798 | .789 | .781 | .773 | .767 | .761 | .756 | .752 | .748 | .745 | .743 | .741 |
| 1.25 | .850 | .841 | .832 | .824 | .817 | .811 | .806 | .801 | .797 | .794 | .791 | .788 |
| 1.50 | .891 | .883 | .875 | .867 | .860 | .854 | .848 | .844 | .839 | .836 | .833 | .830 |
| 1.75 | .923 | .916 | .908 | .902 | .895 | .889 | .884 | .879 | .875 | .871 | .868 | .865 |
| 2.00 | .946 | .940 | .934 | .929 | .923 | .918 | .913 | .908 | .904 | .900 | .897 | .894 |
| 2.25 | .963 | .959 | .954 | .949 | .944 | .940 | .935 | .931 | .927 | .924 | .921 | .918 |
| 2.50 | .975 | .972 | .968 | .964 | .961 | .957 | .953 | .949 | .946 | .943 | .940 | .938 |
| 2.75 | .984 | .981 | .978 | .975 | .972 | .969 | .966 | .963 | .960 | .958 | .955 | .953 |
| 3.00 | .989 | .988 | .986 | .983 | .981 | .979 | .976 | .974 | .971 | .969 | .967 | .965 |

Table 9

Estimated Probability of Consistent Classification ($P_{cc,full}$)

| $\ln[\omega_p]$ | Log Odds Ratio Consistent:Inconsistent, $\ln[\omega_{c/i}]$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Log Odds Passing | .25 | .50 | .75 | 1.00 | 1.25 | 1.50 | 1.75 | 2.00 | 2.25 | 2.50 | 2.75 | 3.00 |
| .00 | .557 | .605 | .647 | .684 | .716 | .745 | .771 | .794 | .815 | .834 | .851 | .866 |
| .25 | .568 | .614 | .654 | .689 | .721 | .749 | .774 | .797 | .817 | .836 | .852 | .868 |
| .50 | .601 | .639 | .674 | .705 | .734 | .759 | .783 | .804 | .823 | .841 | .857 | .872 |
| .75 | .649 | .677 | .704 | .730 | .754 | .776 | .797 | .816 | .833 | .850 | .864 | .878 |
| 1.00 | .705 | .723 | .742 | .761 | .779 | .797 | .815 | .831 | .846 | .861 | .874 | .886 |
| 1.25 | .763 | .772 | .782 | .795 | .808 | .822 | .835 | .849 | .861 | .873 | .885 | .896 |
| 1.50 | .818 | .819 | .823 | .830 | .838 | .847 | .857 | .867 | .878 | .888 | .897 | .906 |
| 1.75 | .864 | .862 | .861 | .863 | .867 | .873 | .879 | .887 | .894 | .902 | .910 | .917 |
| 2.00 | .902 | .897 | .895 | .894 | .895 | .897 | .901 | .905 | .910 | .916 | .922 | .928 |
| 2.25 | .932 | .926 | .922 | .920 | .918 | .919 | .920 | .922 | .926 | .929 | .934 | .938 |
| 2.50 | .953 | .948 | .944 | .941 | .938 | .937 | .937 | .938 | .940 | .942 | .945 | .948 |
| 2.75 | .968 | .964 | .961 | .957 | .955 | .953 | .952 | .951 | .952 | .953 | .955 | .957 |
| 3.00 | .979 | .976 | .973 | .970 | .968 | .965 | .964 | .963 | .963 | .963 | .963 | .965 |

Figure 1

A Bivariate Normal Distribution for Two Half-tests with Cut Scores

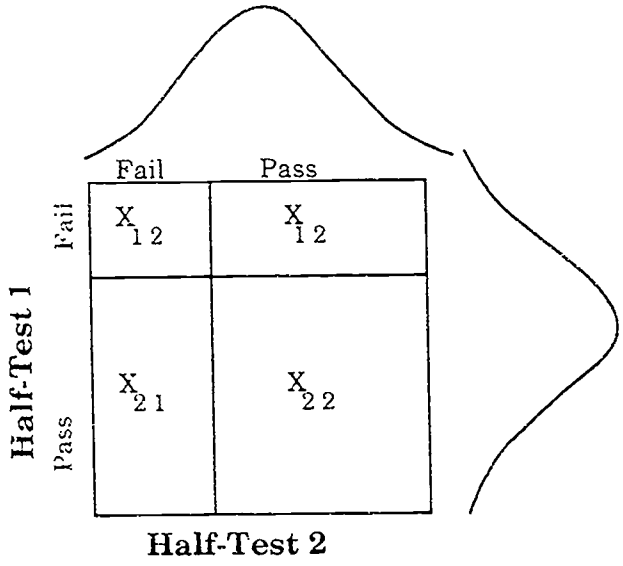

Figure 2

The z score Corresponding to the Average Proportion of Failures

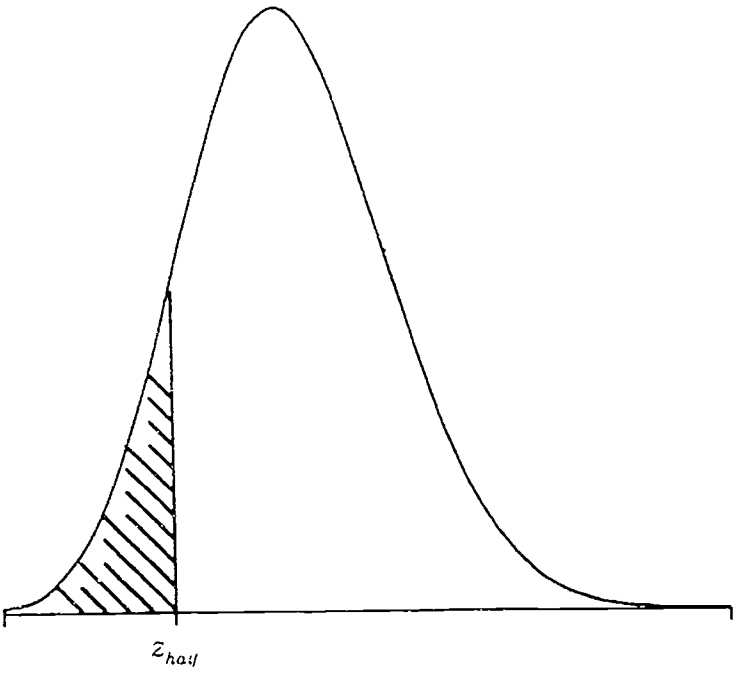

Figure 3

Organization of Some Sample Data

Sample Data

| Half-Test 1 \ Half-Test 2 | Fail | Pass |
|---|---|---|
| Fail | 13 | 12 |
| Pass | 14 | 58 |