DOCUMENT RESUME

ED 273 654 TM 860 503

AUTHOR	Sarvela, Paul D.
TITLE	Discrimination Indices Commonly Used in Military Training Environments: Effects of Departures from Normal Distributions.
PUB DATE	Apr 86
NOTE	36p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Criterion Referenced Tests; *Item Analysis; *Mastery Tests; *Military Training; Postsecondary Education; Raw Scores; Scores; Simulation; Statistical Analysis; *Statistical Distributions; Statistical Studies; Testing Problems; Test Items; Test Theory
IDENTIFIERS	*Discrimination Indices; *Item Discrimination (Tests)

ABSTRACT

Four discrimination indices were compared, using score distributions which were normal, bimodal, and negatively skewed. The score distributions were systematically varied to represent the common circumstances of a military training situation using criterion-referenced mastery tests. Three 20-item tests were administered to 110 simulated subjects. The cutting score on each test was 10 items correct. Three databases were constructed for normal, bimodal, and skewed score distributions. Five item analysis statistics were calculated: the p statistic, two versions of the upper-lower group statistics, the phi coefficient, and the point-biserial correlation. Analysis of variance and t-tests were used to estimate differences between the discrimination index values. With normal data, the second upper-lower statistic produced the largest discrimination values, point-biserial next, and phi coefficient and the first upper-lower produced identical, least discriminating values. Similar results were obtained for the bimodal discrimination indices. The skewed distribution analysis was slightly different, with the first upper-lower results larger than the phi coefficients. The second upper-lower method was not significantly different from the point-biserial correlation. (Suggestions for choosing a method are summarized in a matrix and a decision tree). (GDC)

Discrimination Indices Commonly Used in Military Training Environments: Effects of Departures from Normal Distributions

Paul D. Sarvela, Ph.D.

Ford Aerospace and Communications Corporation *

Western Development Laboratories Division

7100 Standard Drive

Hanover, MD 21076

Paper Presented at

The Annual Meeting of the American Educational Research Association

San Francisco, 1986.

* Paul Sarvela is now with the Dept. of Health Education, College of Education, Southern Illinois University, Carbondale.

Abstract

The unique nature of testing in military training environments (e.g., criterion-referenced testing, bimodal and skewed distributions of test scores) creates special problems in the selection of discrimination statistics used to evaluate test items. This paper describes the findings of a study which compared the results obtained from four discrimination indices when test score distributions were systematically varied (normal, bimodal, and negatively skewed) to represent common military test score distributions. A summary matrix is presented outlining the advantages and disadvantages of each statistic. In addition, a flow chart is included to assist test evaluators make decisions concerning the selection of a discrimination index. The paper concludes with a discussion concerning the practical benefits of each statistic, as well as their relative costs and "ease of use" to the statistically unsophisticated test evaluator.

Discrimination Indices Commonly Used in Military Training Environments: Effects of Departures from Normal Distributions

Introduction

The statistical procedures commonly called item analysis are one way measurement specialists appraise and improve the quality of tests. Of the many forms of item analysis, item difficulty and discrimination indices are most often computed. These statistics provide valuable information concerning the difficulty of the test, as well as the degree to which the test items differentiate between varying achievement levels of students. These data can then be used to increase the reliability of the test (Guilford, 1954).

Although psychometricians working in military training environments recognize the importance of item analysis, they often use criterion-referenced tests (CRTs) to measure student achievement, which frequently produce score distributions that are either bimodal or negatively skewed. Consequently, they work with test score frequency distributions which violate the assumption of normality, an assumption commonly held by many of the "classical" item analysis statistics, such as the upper-lower indices and the point-biserial correlation coefficient. In addition, the small sample size (N = 15 or less) and variability in student achievement found in many military training pilot-study scenarios preclude the application of more sophisticated item analysis strategies, such as the Rasch technique.[1]

Several researchers (e.g., Berk, 1984; Popham, 1981: and Roid & Haladyna, 1982) note that several new CRT-specific item analysis techniques,

---

[1] see Haladyna and Roid (1979) for a discussion concerning Rasch analysis and CRT.

or instructional-sensitivity indices, as described by Haladyna and Roid (1981), have been proposed to address the problems associated with item analysis in CRT testing situations For example, the pre-to-post difference index (PPDI) introduced by Cox and Vargas (1966) and the percentage of possible gain (PPG) developed by Brennan and Stolurow (1971) were both developed to produce item sensitivity indices more appropriate for criterion-referenced testing situations. These indices are excellent methods for obtaining information concerning the quality of CRT items. Unfortunately, they require the test to be administered to students before and after instruction.[2] Often, the military test designer does not have this luxury.

Another CRT approach uses two different groups of students, one group exposed to instruction, while the other group serves as the control (Ellis & Wulfeck, 1982; Popham, 1981). P values are calculated for each group, and the p results from the uninstructed group are subtracted from the instructed group p results, resulting in the discrimination index D_{uigd}. Although this strategy provides solid information concerning the instructional sensitivity of the items, its major disadvantage is that two groups of students are needed, a requirement not always easy to meet in a military testing environment. In addition, the two groups tested must be identical with the exception of treatment, otherwise, variance in the difficulty indices might be attributed to confounding factors outside of the instruction (e.g., one group might be inherently more intelligent than the other group). The problem of randomly assigning students to treatment and control groups might be beyond

---

[2] Popham (1981) notes that an additional disadvantage to these indices is that the pretests might be reactive, and therefore sensitize students to certain items on the pretest

the control of the test specialist, making this method difficult to implement as well.

It is often difficult, if not impossible for the military test specialist to obtain pretest scores, or randomly assign two groups of examinees to instruction or control groups. Therefore, the test designer/evaluator is faced with the problem of maximizing the amount of data concerning test quality that can be gathered from one administration of the test.

One strategy which appears to share the characteristics of both classical techniques and the CRT item sensitivity measures is the Brennan index (Brennan, 1972). This scale is implemented by setting a cut score for mastery on the test, and then dividing the test results into two groups (masters and nonmasters). To obtain BI, the difficulty indices for the nonmasters are subtracted from the indices for the masters (by item). This method is conceptually similar to the upper and lower groups comparison used in classical item analysis (see, for example, Kelley, 1939). The two methods differ in interpretation, however, since one cannot be certain that those in the upper group are truly masters, while those in the lower group are nonmasters. (It should be noted that this same criticism can be applied to the Brennan's technique if the cut score is determined capriciously rather than in a systematic and logical manner.)

When clear-cut mastery or non-mastery cannot be determined (or is to be determined later in the test development process by comparing student performance in the field with their test scores) test specialists must rely on the traditional discrimination indices, despite less than optimum data analysis conditions. Although these statistics and their use have been described in detail by earlier researchers (e.g., Cureton, 1957; Ebel, 1954;

Englehart, 1965; Johnson, 1951) the effects of the violations of the assumption of normality, which commonly occurs in CRT training environments, must be studied in more detail. The purpose of this paper is to describe the findings of a study which compared the results obtained from four different "classical" discrimination indices (two versions of the upper-lower index, r_{pb}, and phi), when test score distributions were systematically varied (normal, bimodal, and negatively skewed) to represent test scores frequently occurring in military testing situations. The paper discusses the practical benefits of each statistic, as well as their "ease of use" to the statistically unsophisticated test evaluator.

Method

Sample and Instrumentation A set of 110 simulated subjects (Ss) were created to represent students enrolled in a military training program. The Ss were "administered" three 20 item tests, scored in a dichotomous manner, with one point assigned to a correct response, and 0 assigned to an incorrect answer. The KR-21 internal consistency reliability index for the normal distribution test was 0.77. The bimodal distribution test KR-21 was 0.87, and the KR-21 coefficient for the skewed test was 0.78. The cut score on each of the tests was set at 10 points.

Procedures Three data bases (normal, bimodal, and skewed) were constructed by varying the distributions of the three sets of simulated test scores. The frequencies of items correct for each S were determined first, dependent on the desired shape of each data base. Next, the item(s) each S answered correctly (1-20) were randomly selected. This randomization produced mean p values of 0.52 for both the normal and bimodal curves. As expected,

the mean p for the skewed distribution was higher (0.74), because more subjects were assigned higher test scores.[3]

The normal curve test score distribution was designed to represent the "control," for which the statistics could be compared, since the majority of psychometric measures commonly used by evaluators require the criterion score variables to be normally distributed. In addition, it represented a common frequency distribution for achievement or aptitude tests used in military settings. In terms of the descriptive statistical properties of the normal distribution data base, the mean was 10.5, with a standard deviation of 4.29. There was a 0.0 value for the skewness coefficient.

The second data base constructed was bimodally distributed. This form of a score distribution is often found in testing situations where there are a group of masters and nonmasters. It also occurs in situations where one group of students receives instruction, while another group does not. This method of studying test items is recommended by Ellis and Wulfeck (1982) in their Handbook for Testing in Navy Schools. The mean score for the bimodal simulation was 10.25. The standard deviation was 5.38, while the skewness coefficient was 0.02.

The third data set (skewed distribution) represented a mastery learning situation. The negatively skewed distribution is commonly found in military environments, where a majority of the students pass the test. These simulation data had a mean of 14.9, and a standard deviation of 3.86. The coefficient of skewness was found to be -0.84, indicating a moderately negatively skewed distribution of the test scores.

---

[3]The 0.52 p value was ideal for the simulation since most authorities recommend a 0.50 p value to study item characteristics (e.g., Kelley, 1939).

A summary of the statistics describing each data base (normal, bimodal, skewed) appears as Table 1.

insert Table 1 about here

Statistical Analyses Five item analysis statistics were calculated in this study: the p statistic, two versions of the upper-lower group statistics (D1 and D2), the phi coefficient, and the point-biserial correlation (r_{pb}).

The difficult index, p, was calculated using the standard formula appearing as equation one:

$$P = \frac{\text{number of correct item responses}}{\text{total number of item responses}} \qquad (1)$$

D1 was obtained by separating those Ss who mastered the learning (masters) from those who failed the test (nonmasters) as suggested by Brennan (1972).[4] (A similar strategy is also used when groups of instructed and uninstructed Ss are available for studying test item characteristics. In this case, one simply substitutes those Ss who received instruction for the masters, and those Ss who did not receive instruction for nonmasters.) In the cases of the normally and bimodally distributed test scores, this strategy was also equivalent to the upper and lower half strategy, because in this study,

[4] mastery was determined by being assigned a test score of 10 or greater

the mastery test score was also the median score for the two data bases. A p value was calculated for each group, and the resulting proportions were subtracted from each other. This statistic is shown as equation two:

$$D1 = \frac{MC}{M} - \frac{NC}{N} \qquad (2)$$

where:

MC = masters who answered correctly

M = total number of masters

NC = nonmasters who answered item correctly

N = total number of non masters

D2 was calculated in a manner similar to D1, however, only the upper and lower 27% test scores were used for the comparisons. An early study by Kelly (1939) demonstrated that this strategy was the most desirable method for studying the effectiveness of items. The method for obtaining D2 appears as equation 3:

$$D2 = \frac{UC}{U} - \frac{LC}{L} \qquad (3)$$

where:

UC = Ss in upper 27% answering correctly

U = total number of Ss in Upper 27%

LC = Ss in lower 27% answering correctly

L = total number of Ss in Lower 27%

Both D1 and D2 have two major assumptions associated with their use:

(1) a normal distribution of criterion scores

(2) equality of mean standard errors of measurement in the upper and lower groups

See Cureton (1957) for a discussion concerning these two assumptions.

The phi coefficient was the third discrimination index used to evaluate the data. In terms of the statistical assumptions associated with the use of the phi coefficient, phi "can be used in any situation in which a measure of the association between two dichotomous variables is desired" (Allen & Yen, p.37, 1979). In this study, the variables were dichotomized by comparing frequencies on each item (pass/fail) with frequencies of test performance (pass/fail). The formula used to obtain the phi values was as follows:

$$\phi = \frac{\chi^2}{n} \qquad (4)$$

where:

n = number of Ss

$$\chi^2 = \sum^{k} \frac{(f_o - F_p)^2}{F_p}$$

where: f_o = observed frequency

F_p = predicted frequency

The point-biserial correlation was the final discrimination statistic used in the study. This statistic was obtained by employing the formula shown in equation 5:

$$r_{pb} = \frac{n\sum X\ Y \quad - \quad \sum X \sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right] \quad \left[n\sum Y^2 - (\sum Y)^2\right]}} \qquad (5)$$

where: X = test item score (0 or 1)

Y = total test score (0 to 20)

N = sample size

There two assumptions most commonly applied to the use of the point-biserial:

(1) a normal distribution of criterion scores should be present

(2) variables should be measured using interval or ratio scales

Analysis of variance (ANOVA) was chosen to estimate significant differences between the various discrimination index values obtained in the

item analysis. The two key assumptions regarding the proper use of ANOVA are (Kachigan, 1982):

(1) The scores in each population are normally distributed

(2) The k population variances are equal (homogeneity of variance)

Upon rejecting the null hypothesis (that the mean values of the item discrimination indices are equal) the paired t-test was applied to the two indices producing the largest average value, to determine if there was a significant difference between the results. The two assumptions for the use of the t-test are:

(1) the scores are normally distributed.

(2) The data are interval in nature

The assumption of normality shared by all tests (with the exception of phi, which is a "distribution free" statistic) was obviously met in the normal distribution data base (see Table 1). Just as obvious, was that the bimodal and skewed data bases violated this assumption. This is not a problem, however, since the central focus of the study was to assess the impact of violations of this assumption.

All data were measured using an interval rating scale. Therefore, the assumption of interval data for analysis was held during the simulation as well.

In terms of the equality of mean standard errors of measurement in the upper and lower groups, since the items correct for each S were randomly assigned, it was concluded that the upper and lower groups would have equal errors of measurement.

With regard to the ANOVA assumptions, the discrimination index values analyzed were somewhat normally distributed, with a slight degree of skewness in the data sets. The skewness coefficients shown in the data are not of

major concern, however, since most authorities agree (see, for example, Games & Klare, 1967) that ANOVA (as well as the T-test) is a robust statistic with regard to violations of the assumption of normality. In terms of the equality of population variances, there do not appear to be significant differences between the item indices studies, therefore, the second assumption was clearly satisfied.

The descriptive characteristics of the variables studied during the ANOVA and t-tests appear as Appendix A.

Results

The results of the analyses for each item appear as Table 2. As expected, the discrimination indices produced different values for different score distributions.

insert Table 2 about here

The normal distribution average p value was 0.52. The mean value for each of the statistics showed clearly that D2 (upper-lower 27%) produced the largest discrimination values, r_{pb} the second largest values, and phi and D1 (upper-lower 50%) produced the least discriminating values. Interestingly, phi and D1 values were identical. ANOVA results suggested that the differences between the discrimination indices were statistically significant between the groups $F(3,76) = 10.5947$, $p < .01$. In addition, the t-test applied to D2 and r_{pb} showed that the differences between these two indices were significant $t(19) = 6.92$, $p < .01$.

insert Table 3 about here

Analysis of the bimodal discrimination indices produces similar results. The average difficulty (p) was 0.52. In addition, D2 produced the largest values, followed by r_{pb}, then phi and D1. Again, the values obtained using phi and D1 were identical. The results of the ANOVA appear as Table 4, showing a significant difference between the 4 groups of indices $F(3,76) = 16.5137$, $p < .01$. The t-test demonstrated that D2 was again superior to r_{pb} for the bimodally distributed test scores $t(19) = 9.88$, $p < .01$.

insert Table 4 about here

The skewed distribution analyses suggested a slightly different pattern. The mean p value for these data was 0.74, clearly showing that more Ss got the items correct than the other two test distributions, an expected finding for a simulation designed to represent a CRT situation. D2 again produced the largest values, followed by r_{pb}. However, in this case, D1 results were larger than those indices obtained using phi. ANOVA results (Table 5) show that there were significant differences between the groups $F(3,76) = 5.4117$, $p<.01$, however, there were no significant differences between D2 and r_{pb}, as suggested by the t-test; $t(19) = 1.04$, p = ns.

insert Table 5 about here

Discussion

The data strongly suggest that the distributions of scores influence the values obtained from the various indices. Clearly, military evaluators should consider the frequency distributions of their test scores when selecting item discrimination indices.

One of the most interesting findings of the study was that the phi coefficient and D1 statistics produced identical values when the test data were bimodally and normally distributed These data suggest that if evaluators are faced with analyzing data with these distributional characteristics, simply calculating the 50% upper-lower index will produce values identical to the phi (provided the cut score happens to be at the median). The evaluator can then use a Pearson r table to estimate the significance of the index, since phi is a special case of r. This strategy can save the evaluator time, because phi is much more difficult to compute than D1.

Another interesting finding was that in the case of the skewed distribution, there were no significant differences between the values produced by D2 and r_{pb}. These results suggest that either method can be used in a skewed distribution setting to obtain essentially the same discrimination values. Therefore, if limited statistical analysis resources are available, the evaluator can use that statistic most easily computed.

Based on the results of this study and the review of the literature, the flow chart appearing as Figure 1 was constructed. Test evaluators can use this flow chart to select that item discrimination statistic most appropriate for their own unique testing situation. The chart begins with the most desirable method for obtaining item instructional sensitivity data. If the conditions cannot be met for the use of this statistic, then, the second most effective statistic is recommended, and so on. (The method of ranking the desirability of the statistics was based on the internal and external threats to validity associated with their use.)

It is important to note that each statistic must be interpreted in its own unique way. An acceptable value for phi might be totally unacceptable for r_{pb}, since each index produces a range of values specific to itself.

For this reason, if quality assurance requirements are placed in the test development product standards, both the statistic and the general level of acceptance should be specified. The problems associated with the violations of the assumptions of normality should also be discussed, outlining which statistic is preferred under a given set of circumstances. This will safeguard both the evaluator and test developer from making inappropriate interpretations of the discrimination statistic values. This recommendation is supported by Englehart (1965) who suggests that critical values for accepting an item's discrimination power are a function of the difficulty of the item.

In terms of the ease of use, costs, and practical benefits of each statistic, the availability of computer resources is a major determining factor in the selection of an item discrimination statistic. Test designers and evaluators who have computer facilities with item analysis programs available can generally disregard the "difficulty" of using various statistics since they are automatically calculated by the computer. However, when dealing with small Ns, where it is not cost-efficient to code and develop a data base, and finally analyze the data, or, where adequate computer facilities are not available, the ease of computation is very important. Undoubtedly, D1 is the easiest of item discrimination indices to obtain. The evaluator must simply rank order the results, divide into upper and lower groups, and compute the results. This method has the added advantage, in the case of normal and bimodal distributions, of being a good estimate of phi. Therefore the significance levels of the indices can be estimated easily. Next easiest is the upper lower 27%. However, a large N should be made available (at least 12 Ss in both the upper and lower groups) otherwise the simple 50% split should be used. Computation of both phi and r_{pb} are more

difficult, and in large N situations, should be employed only through the use of a computer. The statistically unsophisticated evaluator would clearly have more difficulty using these formulae than the simple upper lower groups discrimination index. Table 6 provides a summary matrix of the assumptions, limitations, and ease of use of the statistics described in this study.

insert Table 6 about here

Recommendations for Future Research

Several problems should be investigated in the future to further the knowledge base concerning the use of classical item analysis in CRT settings. One interesting question would be to determine the point where skewness begins to effect the values produced by the discrimination indices. The present study has demonstrated that a moderately skewed distribution produces differences between the statistics that are not found in bimodal and normally distributed test score distributions. A study examining differing degrees of skewness may be needed to help evaluators and researchers select the statistic most appropriate for that level of skewness.

This study employed items with very little variance in p values, by randomly selecting correct responses for each item. Although these results are typical for CRT environments (in that most students get most items correct resulting in a small degree of variance) a study using data with items of differing item variances may reveal different results. This may be an important issue to examine in the future because it is sometimes desirable to use items of differing difficulty values, even in CRT situations.

Finally, the mathematical reasoning behind the equal values for phi

and D1 should be explored, to determine whether the results of this study are a special case of these two statistics (when median and cut scores fall at the same value, and data are normally or bimodally distributed) or whether the mathematical short-cuts derived from the study can be generalized to other data sets as well.

References

Allen, M.J., & Yen, W.M. Introduction to Measurement Theory. Monterey,CA: Brooks/Cole Publishing Co., 1979.

Berk, R.A. (Ed.) A Guide to Criterion-Referenced Test Construction. (2nd ed.) Baltimore, MD: Johns Hopkins University Press, 1984.

Brennan, R.L. A generalized upper-lower item-discrimination index. Educational and Psychological Measurement, 1972, 32: 289-303.

Brennan, R.L., & Stolurow, L.M. An empirical decision process for formative evaluation. Research Memorandum No. 4. Cambridge, MA: Harvard CAI Lab, 1971.

Cox, R.C., & Vargas, J. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Presented at the American Educational Research Association, San Francisco, 1979.

Cureton, E.E. The upper and lower twenty-seven per cent rule. Psychometrika, 1957, 22: 293-296.

Ebel, R.L. Procedures for the analysis of classroom tests. Educational and Psychological Measurement, 1964, 24: 85-90.

Ellis, J.A., & Wulfeck, W.H. Handbook for Testing in Navy Schools. San Diego, CA: Navy Personnel Research and Development Center, 1982.

Englehart, M.D. A comparison of several item discrimination indices. Journal of Educational Measurement, 1965, 2: 69-74.

Games, P.A., & Klare, G.R. Elementary Statistics. New York: McGraw-Hill, 1967.

Guilford, J.P. Psychomtric Methods. (2nd ed.) New York: McGraw-Hill, 1954.

Haladyna, T.M., & Roid, G. The stability of Rasch item and student achievement estimate for a criterion-referenced test. Presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

Haladyna, T.M.. & Roid, G. The role of instructional sensitivity in the empirical review of criterion-referenced test items. Journal of Educational Measurement, 1981, 18, 39-53.

Johnson, A.P. Notes on a suggested index of item validity: the U-L index. Journal of Educational Psychology, 1951, 42: 499-504.

Kachigan, S.K. Multivariate Statistical Analysis: A Conceptual Introduction. New York: Radius Press, 1982.

Kelley, T.L. The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 1939, 30: 17-24.

Popham, W.J. Modern Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1981.

Roid, G., & Haladyna, T.M. A Technology for Test-Item Writing. New York: Academic Press, 1982.

TABLE 1: DESCRIPTIVE STATISTICS: SIMULATED TEST SCORE DISTRIBUTIONS

	Number	Mean	Variance	Std Dev	Std Error	Skewness	Kurtosis
Normal	110	10.5000	18.4174	4.2916	0.4092	0.00000	2.38352
Bimodal	110	10.2545	28.9071	5.3765	0.5126	0.02040	1.59813
Skewed	110	14.9000	14.8982	3.8598	0.3680	-0.84106	3.10766

TABLE 2: ITEM ANALYSIS RESULTS

N = 110

Distribution	Item																					
Normal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	M	SD
p	.45	.62	.53	.54	.45	.51	.54	.58	.47	.46	.53	.55	.56	.53	.58	.57	.53	.50	.48	.49	.52	.05
D1	.22	.30	.47	.38	.38	.29	.27	.26	.47	.38	.29	.55	.18	.33	.33	.57	.25	.42	.20	.40	.35	.11
D2	.40	.50	.56	.53	.73	.30	.53	.37	.67	.60	.56	.86	.37	.46	.66	.70	.44	.46	.37	.43	.53	.14
Phi	.22	.30	.47	.38	.38	.29	.27	.26	.47	.38	.29	.55	.18	.33	.33	.57	.25	.42	.20	.40	.35	.11
r_{pb}	.35	.42	.47	.44	.49	.31	.44	.30	.53	.48	.40	.62	.35	.41	.48	.58	.31	.39	.28	.40	.42	.09
Bimodal																						
p	.49	.49	.55	.50	.49	.55	.47	.54	.54	.51	.52	.54	.55	.54	.48	.49	.48	.50	.50	.52	.52	.03
D1	.56	.45	.48	.44	.49	.55	.53	.48	.62	.46	.51	.33	.42	.62	.44	.49	.55	.47	.33	.58	.49	.08
D2	.77	.70	.57	.54	.60	.63	.73	.54	.84	.57	.67	.56	.73	.73	.57	.60	.63	.57	.57	.73	.64	.09
Phi	.56	.45	.48	.44	.49	.55	.53	.48	.62	.46	.51	.33	.42	.62	.44	.49	.55	.47	.33	.58	.49	.08
r_{pb}	.59	.57	.51	.48	.52	.55	.59	.47	.68	.48	.55	.43	.51	.63	.47	.50	.59	.55	.42	.61	.54	.07
Skewed																						
p	.71	.67	.75	.77	.79	.69	.71	.72	.77	.76	.76	.79	.80	.70	.76	.78	.75	.72	.72	.75	.74	.04
D1	.54	.42	.30	.46	.48	.29	.32	.33	.39	.38	.67	.63	.42	.53	.38	.40	.30	.40	.25	.37	.41	.11
D2	.67	.47	.33	.53	.40	.40	.53	.46	.33	.47	.44	.53	.40	.73	.50	.30	.40	.43	.36	.50	.46	.11
Phi	.42	.31	.24	.39	.42	.23	.25	.26	.33	.32	.56	.55	.37	.41	.32	.34	.24	.31	.20	.30	.34	.10
r_{pb}	.57	.43	.28	.49	.51	.40	.43	.45	.41	.41	.52	.62	.37	.62	.44	.35	.33	.45	.33	.46	.44	.09

TABLE 3: ANOVA: NORMAL DISTRIBUTION

Source of Variation	DF	SS	MS	F-Stat
Among Groups	3	0.4264	0.1421	10.5947
Within Groups	76	1.0195	0.0134	
Total	79	1.4459		

Group Statistics

Group	N	Sum	U-SSQ	Mean	C.V.	S.D.	S.E. (CV)
Norm01	20	6.9400	2.6386	0.3470	31.7361	0.1101	5.5001
Norm02	20	10.5000	5.9084	0.5250	27.4952	0.1443	4.6645
Normphi	20	6.9400	2.6386	0.3470	31.7361	0.1101	5.5001
Normcorr	20	8.4500	3.7329	0.4225	21.9074	0.0926	3.6263

TABLE 4: ANOVA: BIMODAL DISTRIBUTION

Source of Variation	DF	SS	MS	F-Stat
Among Groups	3	0.3106	0.1035	16.5137
Within Group	76	0.4765	0.0063	
Total	79	0.7871		

Group Statistics

Group	N	Sum	U-SSQ	Mean	C.V.	S.D.	S.E. (CV)
BimodD1	20	9.8000	4.9222	0.4900	16.2323	0.0795	2.6333
BimodD2	20	12.8500	8.4041	0.6425	13.7355	0.0883	2.2124
Bimodphi	20	9.8000	4.9222	0.4900	16.2323	0.0795	2.6333
Bimodcorr	20	10.7000	5.8126	0.5350	12.7279	0.0681	2.0448

TABLE 5: ANOVA: NEGATIVELY SKEWED DISTRIBUTION

Source of Variation	DF	SS	MS	F-Stat
Among Groups	3	0.1719	0.0573	5.4173
Within Groups	76	0.8039	0.0106	
Total	79	0.9758		

Group Statistics

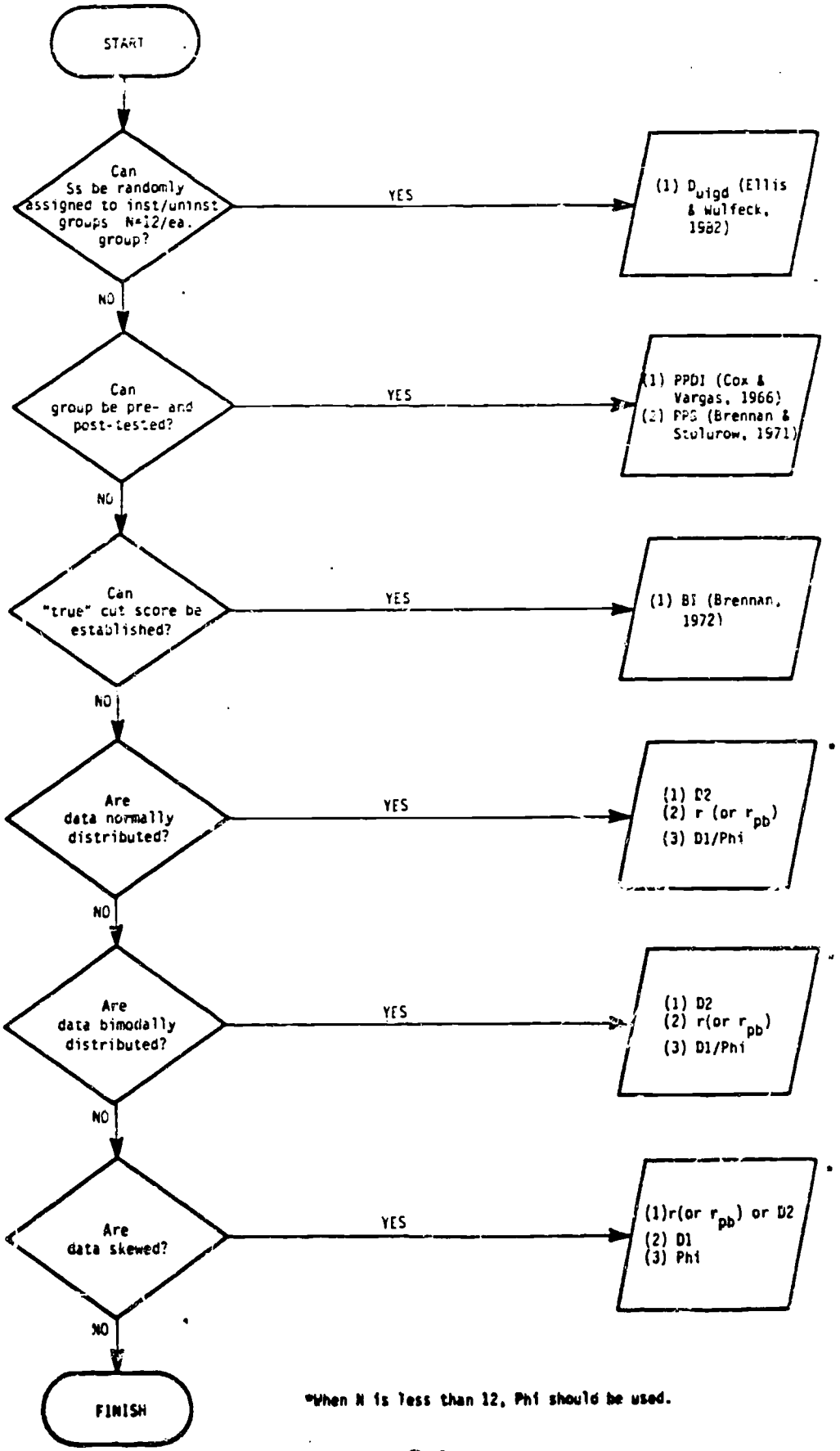
Group	N	Sum	U-SSQ	Mean	C.V.	S.D.	S.E. (CV)
SkewD1	20	8.2600	3.6488	0.4130	27.0665	0.1118	4.5824
SkewD2	20	9.1800	4.4338	0.4590	23.4531	0.1076	3.9069
Skewphi	20	6.7700	2.4757	0.3385	29.0762	0.0984	4.9709
Skewcorr	20	8.8700	4.0961	0.4435	20.8367	0.0924	3.4346

TABLE 6: SUMMARY MATRIX[5]

Measure	Conditions and Assumptions	Calculations	Limitations
PPDI	1. pre & post test Ss 2. N 12 (both groups)	"easy by hand"	must be able to pre and post test
PPG	1. pre & post test Ss 2. N 12 (both groups)	"easy by hand"	must be able to pre and post test
RI	1. master/nonmaster scores 2. N 12 (both groups)	"easy by hand"	must be able to identify masters and nonmasters
D_{uiqd}	1. inst/uninst group sessions 2. N 12 (both groups)	"easy by hand"	must be able to randomly assign Ss to both groups
D1	1. norm distribution 2. = Mean std errors 3. N 12	"easy by hand"	1. assumptions 2. upper group may not be masters
D2	1. norm distribution 2. = Mean std errors 3. N 12	"easy by hand"	1. assumptions 2. upper group may not be masters
r_{pb}	1. norm distribution 2. interval scale 3. N 12	need computer	1. assumptions 2. upper group may not be masters
phi	dichotomous variables	need computer	cut score for pass fail must be set correctly

[5] see Berk (1984) for an excellent discussion on the statistical merits of these statistics.

Figure 1: DISCRIMINATION INDEX SELECTION DECISION TREE



*When N is less than 12, Phi should be used.

APPENDIX A

DESCRIPTIVE STATISTICS: ITEM ANALYSIS INDICES

	Number	Mean	Variance	Std Dev	Std Error	Skewness	Kurtosis
NormP	20	0.5235	0.0022	0.0470	0.0105	0.05111	2.28648
NormD1	20	0.3470	0 0121	0.1101	0.0246	0.46208	2.40361
NormD2	20	0.5250	0.0208	0.1443	0.0323	0.55717	2.65613
NormPhi	20	0.3470	0.0121	0.1101	0.0246	0.46208	2.40361
NormCorr	20	0.4225	0.0086	0.0926	0.0207	0.35805	2.51595
BimodP	20	0.5125	0.0001	0.0269	0.0060	0.12551	1.56193
BimodD1	20	0.4900	0.0063	0.0795	0.0178	-0.31293	2.82427
BimodD2	20	0.6425	0.0078	0.0883	0.0197	0.61979	2.23188
BimodPhi	20	0.4900	0.0063	0.0795	0.0178	-0.31293	2.82427
BimodCorr	20	0.5350	0.0046	0.0681	0.0152	0.21241	2.41453
SkewP	20	0.7435	0.0013	0.0365	0.0082	-0.29532	2.09845
SkewD1	20	0.4130	0.0125	0.1118	0.0250	0.79739	2.99520
SkewD2	20	0.4590	0.0116	0.1076	0.0241	0.88193	3.62875
SkewPhi	20	0.3385	0.0097	0.0984	0.0220	0.83852	3.14578
SkewCorr	20	0.4435	0.0085	0.0924	0.0207	0.35557	2.59345