DOCUMENT RESUME

ED 106 369 TM 004 487

AUTHOR Ascher, Gordon
TITLE Some Nonparametric Approaches to the Use of Criterion-Referenced Statewide Test Results in the Evaluation of Local District Educational Programs.
PUB DATE [Apr 75]
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Academic Achievement; Comparative Testing; Correlation; *Criterion Referenced Tests; Decision Making; Diagnostic Tests; Educational Assessment; Educational Needs; Educational Planning; Hypothesis Testing; Matrices; *Nonparametric Statistics; *Program Evaluation; School Districts; Scores; *State Programs; Statistical Analysis; Testing; *Test Results; Tests of Significance

ABSTRACT

The increased use of criterion-referenced statewide testing programs is an outgrowth of the need for more diagnostic information for planning and decision making than is provided by norm-referenced programs. There remains, however, a need for state agencies to compare the results of local districts to a variety of comparison groups for the purpose of identifying where the greatest needs lie. This paper deals with nonparametric techniques for the comparison of matrices of criterion-referenced scores (rather than the comparison of means). Specific examples include chi square, the median test, rank correlation, the Wilcoxon tests, Kendall's W, and others. (Author)

3.12

SOME NONPARAMETRIC APPROACHES TO THE USE OF CRITERION-REFERENCED STATEWIDE TEST RESULTS IN THE EVALUATION OF LOCAL DISTRICT EDUCATIONAL PROGRAMS

Dr. Gordon Ascher
New Jersey State Department of Education

PRESENTED AT THE 1975 AMERCIAN EDUCATIONAL RESEARCH ASSOCIATION ANNUAL MEETING, WASHINGTON, D.C.

TABLE OF CONTENTS

INTRODUCTION

An increasing number of states have developed criterion-referenced statewide assessment programs and diminished their dependency on norm-referenced programs. The major reason is that CRT provides more information than NRT to educators who use test results for educational planning and decision making. They are more diagnostic. Results may indicate specific strengths and weaknesses in instruction and curriculum. Information is reported for single objectives and clusters of objectives. A single score (as provided by norm-referenced tests) is not as descriptive. However, it provides a handy way to compare groups of students or programs. Program evaluation as well as diagnosis is needed. It is, therefore, necessary for an educator to be able to use criterion-referenced results for both purposes. This paper deals with nonparametric approaches to the problem of program evaluation based on criterion-referenced tests results.

Scope of the Paper. This paper is written for educational practitioners, not statisticians. Its aim is to examine methods which may be applied by those having little formal training in statistics who are willing to use referenced sources. The paper examines the efficacy of using nonparametric tests to compare the criterion-referenced test (CRT) results of two school districts or schools within one district.

Nonparametric Statistics. Siegel (1956, p. vii) presents a concise introduction to nonparametric techniques:

> I believe that the nonparametric techniques of hypothesis testing are uniquely suited to the data of the behavioral sciences. The two alternative names which are frequently given to these tests

> suggest two reasons for their suitability. The tests are often called "distribution-free," one of their primary merits being that they do not assume that the scores under analysis were drawn from a population distributed in a certain way, e.g., from a normally distributed population. Alternatively, many of these tests are identified as "ranking tests," and this title suggests their other principal merit: nonparametric techniques may be used with scores which are not exact in any numerical sense, but which in effect are simply ranks. A third advantage of these techniques, of course, is their computational simplicity. Many believe that researchers and students in the behavioral sciences need to spend more time and reflection in the careful formulation of their research problems and in collecting precise and relevant data. Perhaps they will turn more attention to these pursuits if they are relieved of the necessity of computing statistics which are complicated and time-consuming. A final advantage of the nonparametric tests is their usefulness with small samples, a feature which should be helpful to the researcher collecting pilot study data and to the researcher whose samples must be small because of their very nature (e.g., samples of persons with a rare form of mental illness, or samples of cultures).

THE DATA. Tables 1 and 2 represent the scores achieved by two school districts on a criterion referenced test in mathematics. The test consists of 56 questions (items) clustered under four topics: whole numbers, fractional numbers, measurement and geometry. The result reported for each item is the number of students tested who answered the item correctly and the corresponding percentage of students tested that this number represents. For example, in Table 1 we note that 180 or 75 per cent of the students tested in District A answered item 28 correctly.

The format of these results is realistic. Except that the number of clusters has been reduced, this is the format used in the New Jersey Educational Assessment Program. Similar formats are used in other CRT programs. The numbers reported are

TABLE 1

NUMBER TESTED - 240 STATEWIDE TEST RESULTS DISTRICT A

WHOLE NUMBERS

QUESTION NUMBER	28	1	17	31	11	35	2	18	42	21	55	59	47	56	26	49	34	41
NUMBER CORRECT	180	160	190	160	230	180	100	170	200	110	100	110	100	100	170	110	140	120
PER CENT CORRECT	75	67	79	67	96	75	42	71	83	46	42	46	42	42	71	46	58	50

FRACTIONAL NUMBERS

QUESTION NUMBER	14	37	3	39	57	12	51	62	5	6	8	10	58	48
NUMBER CORRECT	220	160	210	180	100	240	100	190	220	230	220	230	40	190
PER CENT CORRECT	92	67	88	75	42	100	42	79	92	96	92	96	17	79

MEASUREMENT

QUESTION NUMBER	32	50	16	9	45	23	27	7	13	4	46	22	54	20
NUMBER CORRECT	80	80	100	60	50	100	72	63	85	97	43	78	99	110
PER CENT CORRECT	33	33	42	25	21	42	30	26	35	40	18	33	41	46

GEOMETRY

QUESTION NUMBER	36	60	30	44	52	15	40	61	38	25
NUMBER CORRECT	220	220	140	230	160	230	220	210	240	190
PER CENT CORRECT	92	92	58	96	67	96	92	88	100	79

TABLE 2

NUMBER TESTED - 480 STATEWIDE TEST RESULTS DISTRICT B

WHOLE NUMBERS

QUESTION NUMBER	28	1	17	31	11	35	2	18	42	21	55	59	47	56	26	49	34	41
NUMBER CORRECT	360	420	440	440	460	420	440	460	480	142	120	100	170	160	130	170	150	198
PER CENT CORRECT	75	88	92	92	96	88	92	96	100	31	25	21	35	33	27	35	31	41

FRACTIONAL NUMBERS

QUESTION NUMBER	14	37	3	39	57	12	51	62	5	6	8	10	58	48
NUMBER CORRECT	160	160	200	120	100	200	144	124	170	194	86	156	198	220
PER CENT CORRECT	33	33	42	25	21	42	30	26	35	40	18	33	41	46

MEASUREMENT

QUESTION NUMBER	32	50	16	9	45	23	27	7	13	4	46	22	54	20
NUMBER CORRECT	440	320	420	360	200	480	200	380	440	460	440	460	80	380
PER CENT CORRECT	92	67	88	75	42	100	42	79	92	96	92	96	17	79

GEOMETRY

QUESTION NUMBER	36	60	30	44	52	15	40	61	38	25
NUMBER CORRECT	380	480	420	440	460	320	460	280	440	440
PER CENT CORRECT	79	100	88	92	96	67	96	58	92	92

ficticious and contrived for pedagogical reasons.

ANALYSIS

The data in Tables 1 and 2 will be used in an attempt to answer the following questions:

1. Do the total test results indicate significant differences in achievement between the two districts?
2. Do the results indicate that the two districts performed differently on any cluster of test items?
3. Within either district, are there significant differences in performance across clusters?

Three methods of analysis will be described: parametric tests, nonparametric tests for independent samples and non-parametric tests for related samples.

Parametric tests. In most large scale CRT programs individual student results are reported as are the mean and standard deviation of the student results. That is, the CRT program is manipulated to provide some norm-referenced data too. The purpose of the CRT is to provide diagnostic information relative to a large number of objectives and not for gross comparisons between groups. The use of CRT results in a NRT way by some practitioners is a result of their need for both diagnostic and comparative data from the same set of results. We will show here the statistical pitfalls one may encounter if means are used to determine significant differences.

In order to restrict their use in comparisons, the New Jersey program does not report means and variances. Nonetheless, the mean score for a group of students may be calculated from the data in Tables 1 and 2 by summing across all items the number of

students achieving the item. The variance of the students' scores cannot be calculated from these tables, however, so the use of the z or t test is impossible.

It may be argued that in comparing two districts the mean student score is not of interest. Rather, we want to compare the mean number of students achieving each item. This is more in keeping with the use of tests to measure the achievement of stated objectives. We can calculate this mean by summing across all items the number of students achieving each item (as above) and dividing (this time) by the number of items. The variance may be calculated as usual using these same numbers. Table 3 presents summary data for the calculation of means and variances using the number of students achieving each item. It will be noted that the mean for District A is 148.9 and the mean for District B is 297.7. A t test indicates the difference between these means is significant. But this is due to the fact that in District A we tested 240 students and in District B we tested 480. The difference in numbers, of course, would not affect the t test if we used mean student scores.

We still want to use mean numbers of students achieving each item for our comparison but we want to avoid the problem of different size samples. We can do so by calculating the mean percentage of students achieving each item correct. This mean is calculated by summing across all items the percentage of students achieving each item and dividing by the number of items. We can find the mean of the percentages because we have an equal number of students responding to each item. These data are presented in Table 4. Applying the t test results in a t equal to

TABLE 3

MEANS AND STANDARD DEVIATIONS-RAW SCORES

(MEAN NUMBER OF STUDENTS GETTING EACH ITEM CORRECT)

DISTRICT A

	WHOLE NUMBERS	FRACTIONS	MEASUREMENT	GEOMETRY	TOTAL TEST
ΣX	2,630	2,530	1,117	2,060	8,337
ΣX^2	413,100	504,500	94,521	434,000	1,446,121
$\bar{X}$	146.11	180.71	79.79	206.00	148 .88
SD	41.18	60.32	20.38	32.73	61.04
N	18	14	14	10	56

DISTRICT B

	WHOLE NUMBERS	FRACTIONS	MEASUREMENT	GEOMETRY	TOTAL TEST
ΣX	5,260	2,232	5,060	4,120	16,672
ΣX^2	1,923,368	377,584	2,018,000	1,736,000	6,054,952
$\bar{X}$	292.22	159.43	361.43	412.00	297.71
SD	150.74	40.89	120.63	65.46	140.87
N	18	14	14	10	56

TABLE 4

MEANS AND STANDARD DEVIATIONS-PERCENTAGES

(MEAN PERCENTAGE OF STUDENTS GETTING EACH ITEM CORRECT)

DISTRICT A

	WHOLE NUMBERS	FRACTIONS	MEASUREMENT	GEOMETRY	TOTAL TEST
ΣX	1,098	1,057	465	860	3,480
ΣX^2	71,924	87,981	16,383	75,662	251,950
$\bar{X}$	61.00	75.50	33.21	86.00	62.14
SD	17.06	25.08	8.50	13.75	25.47
N	18	14	14	10	56

DISTRICT B

	WHOLE NUMBERS	FRACTIONS	MEASUREMENT	GEOMETRY	TOTAL TEST
ΣX	1,098	465	1,057	860	3,480
ΣX^2	83,874	16,383	87,981	75,662	251,950
$\bar{X}$	61.00	33.20	75.50	86.00	62.14
SD	31.53	8.50	25.08	13.75	25.47
N	18	14	14	10	56

zero since the means are equal. We must conclude that the achievement of District A does not differ from the achievement of District B at least when we analyze the test as a whole. A review of the data, however, indicates that there may be differences between the two districts on several clusters.

For the "Fractions" cluster the t is significant. The same would be true for the "Measurement" cluster. But before we conclude that the t test does work to identify significant differences let us look at the "Geometry" cluster. This t is not significant because, again, the means are equal. We would conclude that the achievement of the two districts on this cluster is equal. A review of the Geometry data in Tables 1 and 2 show that this conclusion is spurious. The two districts differ substantially on this cluster. The t value is an artifact of the data and is a result of treating the cluster results in a gross way. That is, assuming that we could arrange the item results in a random order within each district. This would produce the same mean but certainly not the same meaning since items represent different objectives being assessed.

Before we discount the t test completely, let's examine the results of applying the t test for correlated data to the Geometry cluster results. This would provide some identity to each item. That is, we can compare the results on each item across the two districts. The computation requires us to obtain the difference between A and B's scores on each item. This prevents us from randomizing the scores within each district. Again, the t value is equal to zero. Even though we preserved the identity of each item and prevented the randomization of the

order of items within a district we are still able to randomize the item results as pairs of results and we are still treating the data in a gross fashion. We are seeking, and have not yet found, a statistical test which will provide information about the interaction between items and districts. We want to know if differences in item results within District A are different from differences in item results in District B. We will pursue this.

Nonparametric tests (treating the two districts being compared as independent).

Chi square. Suppose we set up the following table for analysis:

		CLUSTER 1	2	3	4	
District	A	1098	1056	465	860	3480
	B	1098	465	1057	860	3480
		2196	1522	1522	1720	6960

The entry in each cell is the sum across all items of the percentage of students achieving each item (see Table 4). The chi square tests the null hypothesis that the proportions of the row totals assigned to each cell are equal for the two districts and, at the same time, that the proportions of the column totals assigned to each cell are equal for each cluster. That is, it tests whether there are differences in the way students in District A responded to each cluster compared to students in District B.

The computation results in a significant chi square. By observation (or we could apply post hoc contrasts) we note that

the student responses on Clusters 2 and 3 are different in the two districts. This appears to be the test we want. However, if we again review the results on Cluster 4 (Tables 1 and 2) we see that the two districts differ on that cluster but that this chi square masks that difference. This test provides the same problem as one application of the t test above. We conclude that this test "works" only when the numbers of students in each district achieving each cluster differs but it is not sensitive enough for comparisons in which the amount of achievement across all items in the cluster is equal for two districts if the responses differ for items across the cluster.

We can try the chi square with data on each item within a cluster (and we can do this for all of the clusters at once):

ITEM NUMBERS (CLUSTER 4)

		36	60	30	44	52	15	40	61	38	25
District	A	92	92	58	96	67	96	92	88	100	79
	B	79	100	88	92	96	67	96	58	92	92

This chi square is significant. It indicates that there is a difference between the two districts' responses to the cluster because of differences in the districts' responses to each item.

There is one major problem with using the chi square test in this way; it is overly sensitive to differences in single

items. That is, if we observed the following data:

ITEM NUMBERS

	36	60	30	44	52	15	40	61	38	25
A	80	80	80	80	80	80	80	80	80	80
B	10	80	80	80	80	80	80	80	80	80

and applied the chi square test we would find a significant chi square. We are "safe" if we do not accept the results as evidence of the difference between the two districts on the cluster as a whole but go back and look for the source of the difference. This point is essential.

Other tests which assume the two districts to be independent fail to provide evidence of some differences (i.e., Cluster 4 type differences) because these tests examine the total distribution. One illustration is presented.

The Mann-Whitney U Test is used to determine whether two groups of "scores" have been drawn from the same population. It is not sensitive to different arrangements of scores within the two distributions. This point concerns us here and, therefore, this test and others making this assumption (e.g., median test, Kolmogorov-Smirnov two-sample test, Wald-Wolfowitz runs test) provide spurious results.

To apply the Mann-Whitney U test we rank all of the scores (for both districts) while tagging each score so that we remember which district it represents.

SCORE	DISTRICT	SCORE	DISTRICT
58	A	92	A
58	B	92	B
67	A	92	A
67	B	92	B
79	A	96	A
79	B	96	B
88	A	96	A
88	B	96	B
92	A	100	A
92	B	100	B

We calculate U by noting how many A scores preceed each B score and sum these. The first B score is preceeded by one A score. The second B score is preceeded by 2 A scores. The third B score is preceeded by 3 A scores and so on: U= 1+2+3+4+5+6+7+8+9+10= 55. We cannot reject the null hypothesis and must conclude that the two districts are equal on this cluster which we know is not so.

Again, the reason for this spurious result is that the test looks at two sets of numbers which can be arranged randomly within each sample (district) and is not sensitive to two different arrangements of the same numbers. We still seek a test which is sensitive to differences of this kind but is not overly sensitive as was the chi square above. It occurs to us to use a test which treats the results of each district on each item as related. But this would involve tests usually used with related samples. Can we argue convincingly that these districts (or even two schools within a district) are related? Or need we? If the tests can be adapted for our kind of data we can use them.

Nonparametric tests (those which treat the data as drawn from related samples).

Wilcoxen matched-pairs signed-ranks test. An example: Suppose we collected the following scores on 10 people each of which had been exposed to two treatments (in random order) and tested after each treatment

Subject	Treatment 1	Treatment 2	d	Rank of d	Rank with less frequent sign
A	36	37	-1	-1	1
B	30	16	14	7	
C	47	40	7	4	
D	56	38	18	9	
E	14	20	-4	-2	2
F	42	19	23	10	
G	21	10	11	6	
H	25	9	16	8	
I	48	40	8	5	
J	24	19	5	3	
					T = 3

We observe a T of 3 which is significant and tells us that the two samples are drawn from different populations. Since this test involves comparing pairs of data and is sensitive not only to differences but to the magnitude of the differences it may be appropriate to our needs. We can use our items in place of subjects and our districts in place of treatment. No assumption is made about the relationship between the two districts. This is only an attempt to utilize existing tests with known sampling distributions for our purposes.

Using the data from Cluster 4:

Item	District A	District B	d	Rank of D
36	92	79	13	5.5
60	92	100	-8	-3.5
30	58	88	-30	-9.5
44	96	92	4	1.5
52	67	96	-29	-7.5
15	96	67	29	7.5
40	92	96	-4	-1.5
61	88	58	30	9.5
38	100	92	8	3.5
25	79	92	-13	-5.5

T = 27.5

This result is not significant and we must conclude that the two districts scored in a similar way, which we know is not true. However, we must remember that this is a test of the central tendency of the two distributions. To illustrate, consider these two sets of scores:

Item	A	B
A	1	5
B	2	4
C	3	3
D	4	2
E	5	1

We need not perform the computations to see that the central tendency of both sets of scores is the same, i.e., the median score in both distributions is 3. Yet, the responses to each of the items in the two groups is quite different. Perhaps we can use a correlational technique.

Nonparametric tests (correlational techniques)

Spearman rank correlation. The Spearman rank correlation

coefficient for the data above is -1. This indicates a high degree of relationship in an inverse fashion. If these were the scores of Cluster 4 for both districts the Spearman test would tell us that the scores on individual items do vary a great deal between the two districts.

The Spearman rho calculated for the following data is +1 indicating a high correspondence on each item:

Item	A	B
A	1	1
B	2	2
C	3	3
D	4	4
E	5	5

The Spearman rho for the following data is equal to zero indicating no systematic relationship item by item:

Item	A	B
A	1	2
B	2	5
C	3	3
D	4	1
E	5	4

Yet we see there are differences across items especially when we remember that each item represents a distinct educational objective. We can adapt the Spearman rho to our needs by changing the null hypothesis and the rejection region. Simply put, any value of rho which is not both positive and significant may be regarded as an indication of differences between the two sets of scores being compared worthy of further examination.

Consider the data in Cluster 4. The Spearman rho is -.03. Since this is not positive and significant we may interpret it as an indication of differences between the two districts. We can compare the results of more than two districts (or, for example, the results of many schools in a district) at one time by applying Kendall's coefficient of concordance test.

SUMMARY

- We are seeking a test of significance for comparing the results of two groups on a criterion referenced test.
- We wish to avoid measures of central tendency.
- We wish to preserve the information provided by each test item since items represent educational objectives.
- Some success is found with the use of chi square.
- It is recommended that correlational techniques be used with an adjusted null hypothesis and rejection region.
- Because of sample size, nonparametric correlational techniques are recommended.

REFERENCES

Ascher, G. Utilizing Assessment Information in Educational Planning and Decision-Making. Trenton, N.J.: State Department of Education, 1973.

Bradley, J.V. Distribution-Free Statistical Tests. Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Conover, W.J. Practical Nonparametric Statistics. New York: Wiley, 1971.

Pierce, A. Fundamentals of Nonparametric Statistics. Belmont, Ca.: Dickenson, 1970.

Tate, M.W. and Clelland, R.C. Nonparametric and Shortcut Statistics. Danville, Ill.: Interstate, 1957.

Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.