DOCUMENT RESUME

ED 287 859 TM 870 560

AUTHOR Johnstone, Whitcomb G.; Wilson, Michael J.
TITLE Standardized Test Selection Practices in the Public Schools.
PUB DATE Apr 87
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, DC, April 21-23, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Testing; Elementary Secondary Education; *Evaluation Criteria; Information Utilization; National Surveys; *Psychometrics; *Public Schools; Rating Scales; School Districts; Standardized Tests; *Test Format; Testing Programs; *Test Selection
IDENTIFIERS Test Directors

ABSTRACT

This survey investigated: (1) the extent that psychometric criteria are used in test selection; (2) the weight given to different types of psychometric information relative to other qualities of the test; and (3) possible reasons for differences among districts. Directors of testing offices, including members of the National Association of Test Directors, from 200 school districts were asked to rate 13 areas considered to be evaluative criteria in test selection. Component items within each global area were also rated. Eighty-one districts responded. Analysis showed that importance was attached to test validity, test reliability, and norming/standardization. Test administration was the only non-psychometric criterion ranked as high as any psychometric criterion. In general, it was found that school districts tended to place heaviest emphasis on psychometric criteria in evaluating tests, though certain non-psychometric criteria involving directions given to students and examiners were also weighted highly. The relative importance of psychometric and non-psychometric criteria in test selection did not appear to vary with district size or other characteristics, although the level of importance assigned to either type of criteria appeared to correlate with the level of importance assigned to the other. (MDE)

Standardized Test Selection Practices In the Public Schools

Whitcomb G. Johnstone

Irving (TX) Independent School District

and

Michael J. Wilson

North Texas State University

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D. C., April, 1987.

Abstract

Standardized Test Selection Practices
in the Public Schools

Whitcomb G. Johnstone
Irving (TX) Independent School District

Michael J. Wilson
North Texas State University

The objectives of this survey were to determine the extent that psychometric criteria are used in standardized test selection, the weight given to different types of psychometric information, and the importance of psychometric information relative to other qualities of a test. The survey was administered to a sample of directors of testing offices identified through the membership list of the National Association of Test Directors and other means. The results indicate that traditional psychometric criteria such as content validity, test reliability and norming/standardization appear to be the most important to districts. Clarity of directions to students and examiners stand out among non-psychometric criteria for test evaluation. Because publishers must respond to the marketplace to survive it is important for practitioners to know what signals publishers are receiving from the marketplace and for publishers to know what the users of standardized tests want. Test selection practices will shape future standardized tests.

Standardized Test Selection Practices in the Public Schools

The selection of standardized achievement tests is one of the most common and most important activities of school district research and testing offices (Wilkens, 1981; Hrul and Casserly, 1982). Iwanicki (1980), for one, has summarized many of the features of recent standardized achievement tests that schools should take into account in selecting a test. While prescriptions for how to select a standardized achievement test abound (Strozeski and Mason, 1986; Mehrens, 1984; Perlman, Junker and Rice, 1984; Messick, 1981; Petrosko and Shani, 1977; Ward, Blackman, Hall and Mazur, 1974) there is little information about how districts actually go about making the decision to adopt a test. This lack of information is unfortunate because, as one representative of a major publisher stated in a recent symposium (Drahozal, 1986), "Test selection procedures impact the nature of the tests." This paper reports the results of a survey intended to assess the importance of psychometric and non-psychometric criteria in school district test selection practices. The objectives were:

1. To determine the level of use of types of psychometric and non-psychometric information in the selection of a standardized norm-referenced achievement test.
2. To determine the relative weight given the different types of information in the selection process.
3. To determine possible reasons for differences among districts.

PROCEDURES

Instrument

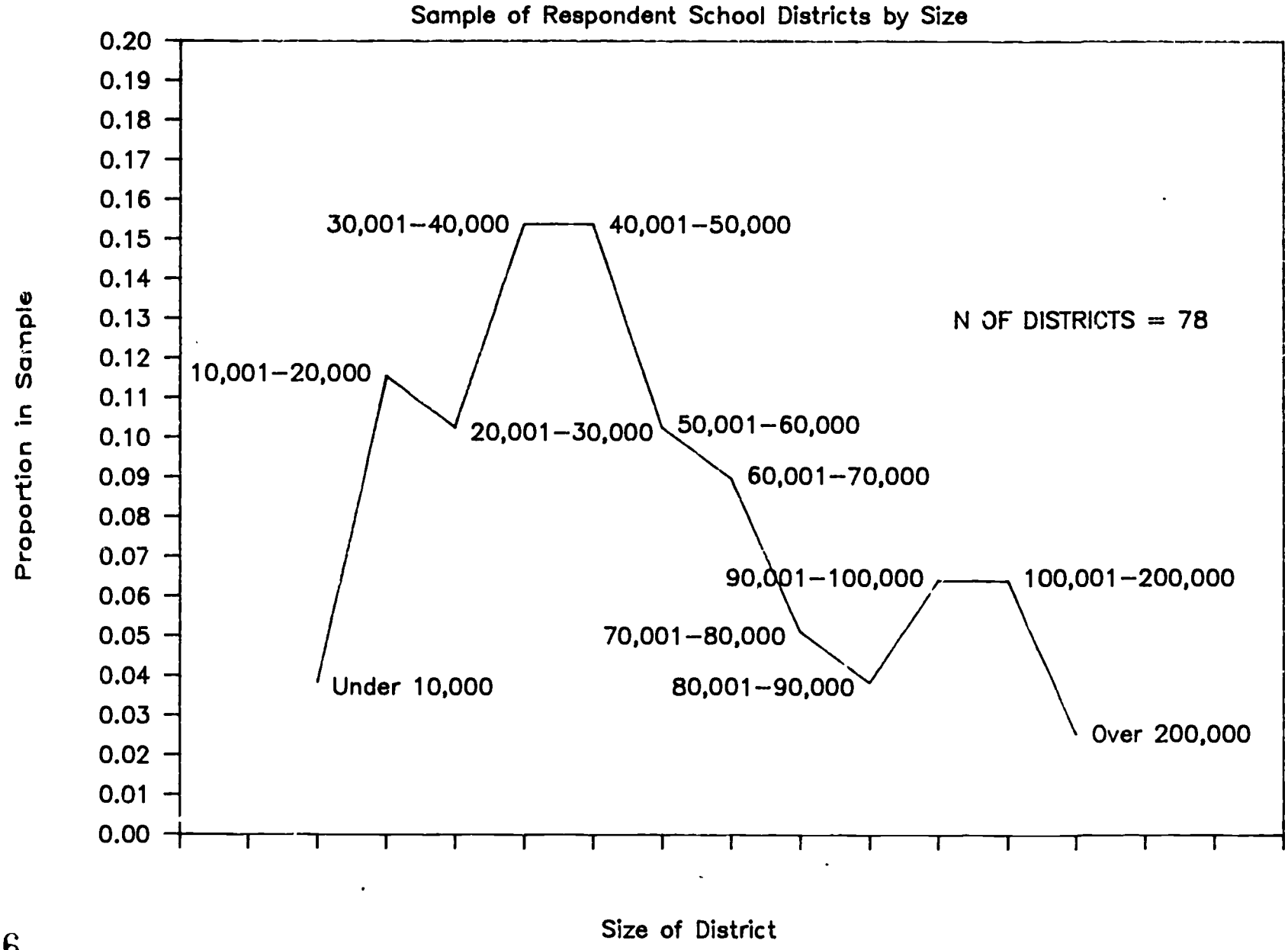
The questionnaire was developed through discussions with test directors who had recently completed test adoptions. The majority of items were framed as five-point Likert scales anchored by the descriptors "not important" for a rating of one and "highly important" for a rating of five. From four to sixteen component items were developed in each of thirteen global areas considered to be evaluative criteria in test selection. Respondents were asked to rate all of the global criteria for importance, then to rate the component criteria within each area. Some of these criteria related to psychometric characteristics and some did not as shown below:

Psychometric	Non-Psychometric
Validity	Administration
Item Analysis	Costs
Reliability	Scoring/Score Reporting
Norming/Standardization	Publisher Services
Bias	Functional Testing
Equating of Forms	Test Appearance
	Test Design

Sample

The survey sampling frame was developed in several ways. One part was established by acquiring the list of members of the National Association of Test Directors. Another part of the frame was acquired by calling test publishers for lists of districts which had recently selected a standardized test. Finally, a small part included school districts personally known to the investigators. The frame developed

Figure 1


Sample of Respondent School Districts by Size
Proportion in Sample
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30,001–40,000
40,001–50,000
N OF DISTRICTS = 78
10,001–20,000
20,001–30,000
50,001–60,000
60,001–70,000
90,001–100,000
100,001–200,000
70,001–80,000
Under 10,000
80,001–90,000
Over 200,000
Size of District

from these sources was fairly small, somewhat over 200 public school districts, but quite diverse in size and geography.

Since the sampling frame was relatively small it was determined to mail a questionnaire to all of the districts in the frame. When possible, telephone contact was established with the district before hand to facilitate participation, determine the proper recipient and establish rapport. Of slightly over 200 questionnaires distributed, 81 were returned. The distribution of respondent districts by size is depicted in Figure 1.

Analysis

Descriptive statistics for the ratings of the global and component criteria were generated and cross tabulations by district size, technical expertise, length of the selection process and the number of groups or committees involved were prepared.

RESULTS

Results are presented as they relate to the three objectives for the research stated in the introduction.

Objective 1 Determine the level of use of psychometric and non-psychometric information in the selection of standardized tests.

Table 1 presents an item analysis of the importance ratings for the global test evaluation criteria in the selection process. Presuming that districts use most often the criteria that they regard as most important, districts appear to emphasize the traditional areas of validity, reliability and norming/standardization.

The individual component criteria within these three broad areas that were rated as significantly more important than the average

rating for all items within the same global category are presented in Table 2. Significance was established as falling beyond the upper bound of the 95 percent confidence interval about the mean rating for all items.

Table 1

Major Areas of Evaluation for Test Selection Decisions By Levels of Importance in Percent

Major Areas	Missing	Level of Importance Lowest 1	2	3	4	Highest 5
Validity	0.0	0.0	0.0	2.5	17.3	80.5
Reliability	0.0	0.0	0.0	1.2	19.8	79.0
Norming/Standardization	2.5	0.0	0.0	4.9	23.5	69.1
Bias	0.0	1.2	1.2	16.0	32.1	49.4
Item Level Analysis	0.0	2.5	6.2	13.6	33.3	44.4
Administration	0.0	3.7	9.9	18.5	40 7	27.2
Equating Forms	2.5	2.5	3.7	17.3	49.4	24.7
Test Design	6.2	2.5	3.7	19.8	40.7	27.2
Scoring/Score Reporting	1.2	14.8	9.9	17.3	22.2	34.6
Costs	0.0	4.9	11.1	34.6	29.6	19.8
Test Appearance	1.2	3.7	7.4	38.3	34.6	14.8
Publisher Services	0.0	11.1	18.5	33.3	18.5	18.5
Functional Testing	4.9	9.9	14.8	33.3	25.9	11.1

It is interesting to note under the criteria for validity that the mean rating "Match with State Curriculum Guidelines" was significant in the high importance direction while the mean rating for "Match with District Curriculum Guidelines" (not shown in Table 2) was significant in the low importance direction.

Objective 2 Determine the relative weight given the different types of information in the selection process and possible reasons for differences among districts.

Figure 2 was prepared for the mean importance ratings given to the 13 global of criteria for standardized test evaluation in the

Table 2

Specific Criteria Rated Significantly Higher in Importance Than Others Within the Areas of Validity, Reliability and Norming/Standardization.

Criteria	Mean Rating
Validity	
Adequate Sampling of Achievement Content	4.78
Relevance of Items to Student Experience	4.57
Match with State Curriculum Guidelines	4.40
Reliability	
Test-Retest Reliability	4.44
Internal Consistency	4.40
Standard Error of Measurement	4.35
Subtest Reliabilities	4.32
Norming/Standardization	
Percentile Scores Provided	4.64
Evidence of National Representation	4.63
Fall & Spring Norm Availability	4.62
Standard Scores Provided	4.42
Age of Available Norms	4.37
Representation of Similar Populations	4.27
Publisher Adherence to Sampling Plan	4.26

selection process. Psychometric criteria clearly stand out as the most important. Of the six psychometric criteria only one, forms equating, was rated lower than the highest ranked of the non-psychometric criteria. Validity and reliability were ranked first among the psychometric criteria, followed by norming/standardization, bias studies and item analysis in order.

Among the non-psychometric criteria, test administration was rated most important, somewhat higher than forms equating. Test design was next. Scoring/reporting, costs and appearance were all rated about equally at the next lower level of importance. Publisher service and functional level testing were rated the least important.

Participants not only rated the thirteen global areas of test evaluation directly, but within each area they rated several specific possible component criteria. Figure 3 depicts the relationship between the thirteen mean global ratings and mean ratings over the specific component criteria within each of the evaluative areas. It shows that these two ways of viewing the relative importance of the test selection criteria support each other. That is, respondents who rated global areas as higher in importance tended to also rate the component criteria in those areas higher than the individual criteria in other areas.

Two areas appear to somewhat contradict this general finding. Both test appearance and test design would be viewed as relatively more important based on the mean ratings of the component criteria for these areas than was apparent for the global ratings.

Figure 2

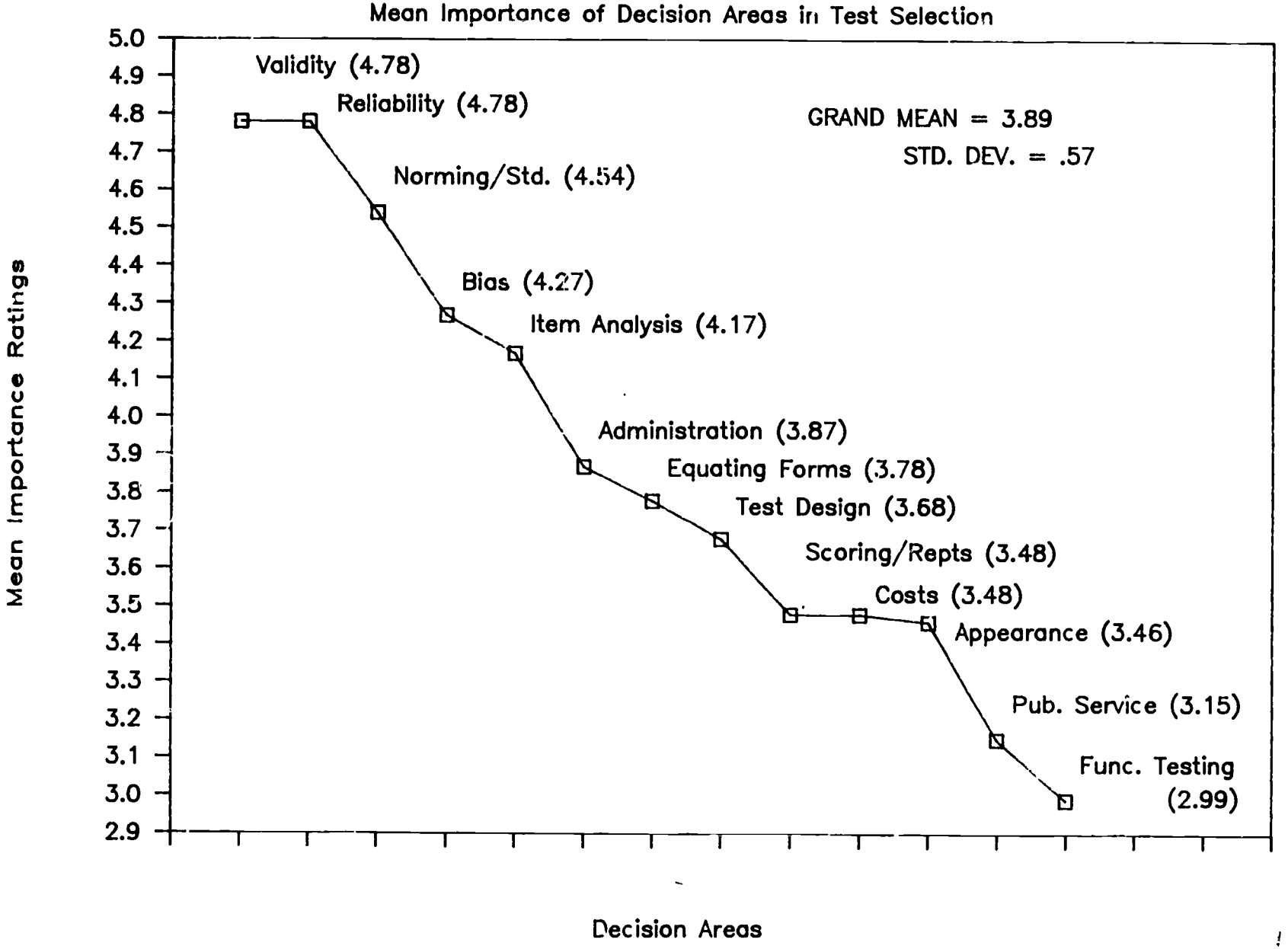
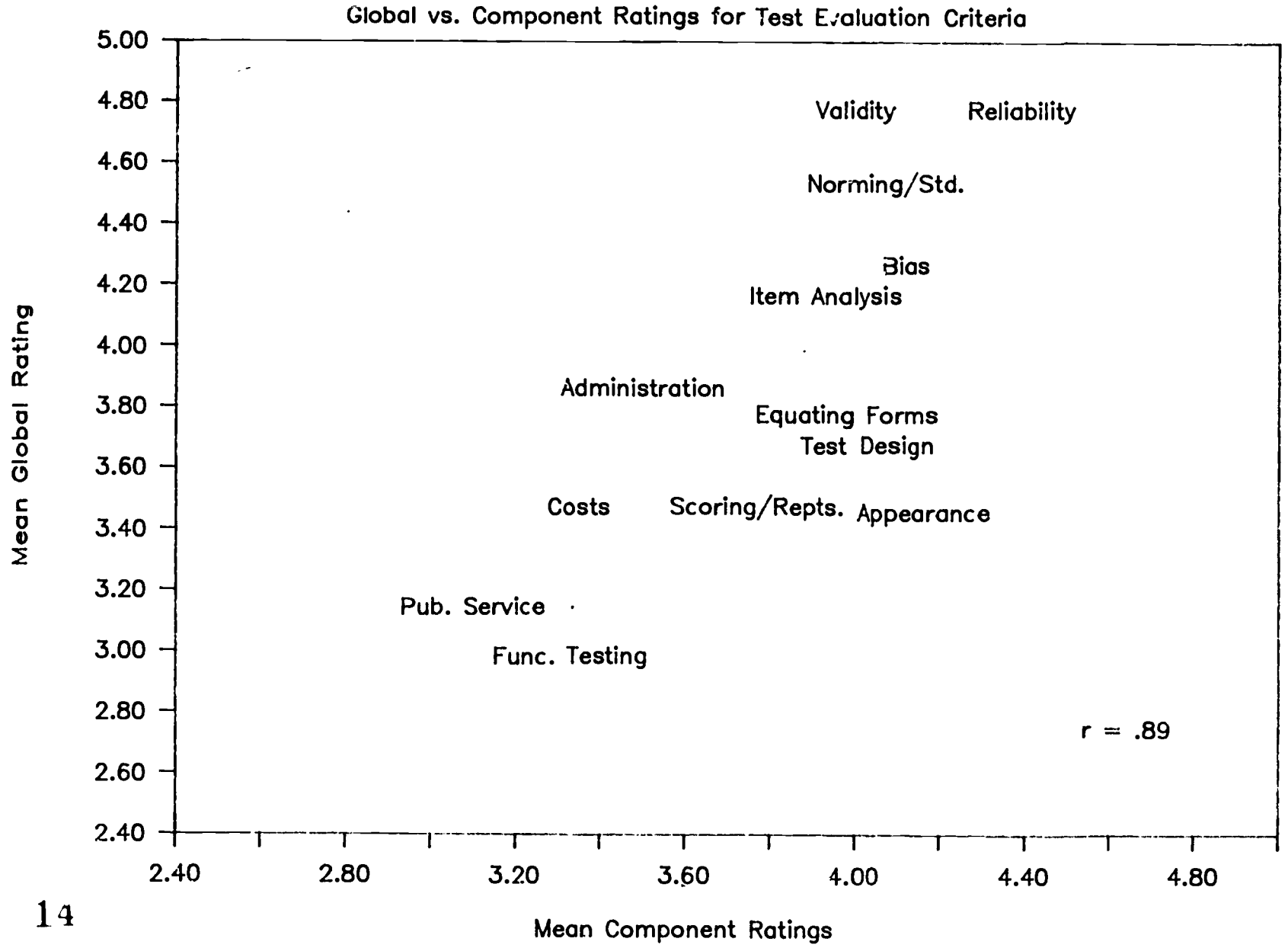

Mean Importance of Decision Areas in Test Selection
Mean Importance Ratings
5.0
4.9
4.8
4.7
4.6
4.5
4.4
4.3
4.2
4.1
4.0
3.9
3.8
3.7
3.6
3.5
3.4
3.3
3.2
3.1
3.0
2.9
Validity (4.78)
Reliability (4.78)
Norming/Std. (4.54)
Bias (4.27)
Item Analysis (4.17)
Administration (3.87)
Equating Forms (3.78)
Test Design (3.68)
Scoring/Repts (3.48)
Costs (3.48)
Appearance (3.46)
Pub. Service (3.15)
Func. Testing (2.99)
GRAND MEAN = 3.89
STD. DEV. = .57
Decision Areas

Figure 3


Global vs. Component Ratings for Test Evaluation Criteria
Mean Global Rating
5.00
4.80
4.60
4.40
4.20
4.00
3.80
3.60
3.40
3.20
3.00
2.80
2.60
2.40
Validity
Reliability
Norming/Std.
Bias
Item Analysis
Administration
Equating Forms
Test Design
Costs
Scoring/Repts.
Appearance
Pub. Service
Func. Testing
r = .89
2.40
2.80
3.20
3.60
4.00
4.40
4.80
Mean Component Ratings

This apparent contradiction is explored in Table 3, which presents the ten component criteria with the highest mean ratings for importance across all thirteen global areas of test evaluation.

Table 3

Ten Highest Mean Ratings of Importance
of Psychometric and Non-psychometric Characteristics

Evaluation Area	Criterion	Means
Validity	Adequate Sampling of Achievement Content	4.78
Test Design	Administration Directions	4.65
Test Design	Directions to Students	4.64
Norming	Percentile Scores Provided	4.64
Item Analysis	Readability	4.63
Norming	Evidence of National Representation	4.63
Item Analysis	Item Clarity	4.62
Norming	Availability of Spring & Fall Norms	4.62
Validity	Relevance of Items to Students Experience	4.57
Appearance	Appropriateness of Printed Directions	4.56

On the one hand, items related to test design and appearance were among the most important individual criteria, even though as global areas they ranked low in importance. On the other hand, no single reliability criterion appeared among the most important component criteria even though the global area of reliability was rated as most important along with validity.

Objective 3: Determine possible reasons for differences among districts in the application of test selection criteria.

It was supposed that some differences in the importance of psychometric and non-psychometric criteria in evaluating tests could be accounted for by such district characteristics as size, the type of department conducting the test selection, the personnel involved and the presence of a state selection. A few of the component criteria were found to have a weak relationship with the size of the district. These are shown in Table 4. Otherwise, no significant relationships were found for either individual component criteria or for the global criteria.

Table 4

Test Evaluation Criteria Most Likely to Relate to Size of District

Relationship	Significance Level	Size *
Publisher scoring	.01	small
District option to score test	.05	large
Scoring program available for 3rd party mainframe use.	.05	middle

* Size of district with highest importance

The importance of non-psychometric criteria was explored for varying levels of importance assigned by districts to the psychometric criteria. Table 5 shows that significant relationships exist between levels of importance assigned to certain non-psychometric criteria and the overall importance assigned to psychometric criteria. The

strongest relationships were observed for the global areas of functional testing and test design.

Table 5

Mean Levels of the Importance of Non-psychometric Criteria Across Levels of the Importance of psychometric Criteria

Non-Psychometric Characteristics	Psychometric Characteristics Low	Low Medm.	High Medm.	High	ETA
1-Administration of Test *	3.94	3.45	3.95	4.37	.37
2-Cost	3.68	3.05	3.62	3 74	.26
3-Scoring	3.16	3.42	3.19	4.26	.31
4-Services *	2.68	2.90	3.62	3 47	.32
5-Func. Testing***	2.50	2.67	3.19	4.05	.53
6-Test Appearance	3.21	3.21	3.81	3.79	.30
7-Test Design ***	3.47	4.24	4.24	4.42	.47

Note: *.05 ***.001

CONCLUSIONS

Objective 1 What is the level of importance of psychometric and non-psychometric information in the selection of standardized achievement tests?

The ranked order of the global test selection criteria in Figure 2 indicates that the psychometric criteria are the most important consideration in the selection of achievement tests. Test administration is the only non-psychometric criterion which is ranked as high as any psychometric criterion. When the same thirteen criteria are ranked using the composite means of the individual

component criteria, there is a correlation of r=.89 between the component criteria means and the means for the global rating as shown in Figure 3. Using the composite component criteria means, appearance and test design replace administration as the highest ranked non-psychometric selection criteria, but psychometric criteria still appear to be the most important in the selection of achievement tests.

Table 2 seems to support many of the orthodoxies that are commonly promoted as important in achievement test selection. For example, the importance of content validity, the necessity of interpreting responses against the norm group, and statistical and content considerations of bias. There is also evidence of a practical concern in the importance of matching student experience and state curriculum guidelines with test content.

The non-psychometric evaluation criteria are primarily practical concerns. Following is a list of the ten most important non-psychometric selection criteria.

Test administration directions
Student directions
Appropriateness of printed directions
Appropriateness of print size
Clarity of examples
Directions for responding to student questions
Age appropriateness of illustrations
Test reports for teachers
Timely return of score reports
Quality of graphics

It can be seen that all but two of the concerns are directly involved with the administration of the test itself. The two that are not are related to scores.

In the list of the ten highest rated individual criteria in Table 3, the importance of practical versus technical is even more apparent. For example, directions, readability, item clarity and relevance of items appear to be as important as the technical concerns of sampling of content, percentile scores, and norming samples.

Objective 2 What is the relative importance of psychometric and non-psychometric test selection criteria?

A general pattern seems to indicate that those districts which rated psychometric criteria high also rated the non-psychometric criteria high. From Table 5, two of the non-psychometric criteria appear to be clearly different, functional testing and test design. Two more, test administration and publisher services, may also be different. Table 6 below provides a rank order placement for non-psychometric criteria within each level of psychometric importance.

In the table, scoring and cost seem to be the most erratic, while test design and test administration seem to be the most stable in relation to all of the other non-psychometric criteria. Test design and administration seem to be the most important criteria across all levels of psychometric importance; publisher services and functional testing seem to be the least important.

Objective three What are some possible reasons for the differences among districts?

Very little was found in the analysis that could be considered of an explanatory nature. Most relations that were found were weak. It can be seen in Table 4 that the smaller districts were more interested in assistance with technical details than the larger ones. The larger districts appeared to find the use of their mainframe computers to

score and interpret the tests themselves more important than the smaller ones. This is probably true of the larger districts because the smaller ones do not have the hardware nor the technical expertise to accomplish such a task.

Table 6

Ranking of Non-psychometric Concerns Across Levels of Psychometric Importance

Rank	Low	Low-med.	High-med.	High
1st	Admin.	Admin.	Admin.	Admin.
2nd	Cost	Scoring	Design	Scoring
3rd	Design	Design	Appearance	Design
4th	Scoring	Cost	Cost	Func.Tstng.
5th	Appearance	Appearance	Services	Appearance
6th	Services	Services	Func.Tstng.	Cost
7th	Func.Tstng.	Func.Tstng.	Scoring	Services

SUMMARY

In general, it was found that school districts tend to place heaviest emphasis on psychometric criteria in evaluating tests, though certain non-psychometric criteria related to directions given to students and examiners are also weighted highly. The relative importance of psychometric and non-psychometric criteria in test selection does not appear to vary with district size or other characteristics measured on the survey, although the level of importance assigned to either type of criteria appears to correlate with the level of importance assigned the other.

Most educators can agree that the standardized achievement test has a significant place in the schools. Information generated from such tests is used to evaluate instruction, provide accountability, benchmark student progress, communicate with parents and school boards and to make important decisions about students. If only 70% of all U.S. public school students took a standardized test it is estimated that $40 million would be spent annually. Given the size of the market and the impact of standardized tests on school districts, the process of selecting tests has received surprisingly little study. Although many authors have written about test selection, only one empirical study of a district selection process (Perlman, et al., 1984) was found in a search of the literature. Publishers must respond to the marketplace to survive. It is important for practitioners to know what signals publishers are receiving from the marketplace and for publishers to know the basis on which users select standardized achievement tests. The present investigation is an attempt to generate a better understanding of the process of test selection by looking at the importance assigned to test selection criteria of different kinds by school districts across the country.

REFERENCES

Drahozal, Edward C. Readability, Correlations/Cross-Referencing to State Learner Objective and Other Important and Unimportant Considerations in the Achievement Test Selection Process. Symposium Presented at the Annual Meeting of the Southwest Educational Research Association, Houston, Texas, February, 1986.

Hrul, Judith and Casserly;y, Michael. Student Performance Assessment in the Great City Schools. Unpublished Paper of the Council of the Great City Schools, 1982.

Iwanicki, Edward F. A new generation of standardized achievement test batteries: A profile of their major features. Journal of Education Measurement, 17, 155-162, 1980.

Mehrens, William A. National tests and local curriculum: Match or mismatch? Educational Measurement: Issues and Practices, 3, 9-15, 1984.

Messick, Samuel. Evidence and ethics in the evaluation of tests. Educational Researcher, 10, 9-29, 1981

Perlman, Carole L., Junker, Linda K. and Rice, William K., Jr. Choosing an Achievement Battery: A Case Study. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, 1984.

Petrosko, Joseph M. and Shani, Esther. Structural Components Revealed by Evaluating the Quality of Elementary School Tests. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1977.

Strozeski, Michael and Mason, Barbara. Materials to Accompany Session 33: Readability, Correlations/Cross-Referencing to State Learner Objectives, and Other Important and Unimportant Considerations in the Achievement Test Selection Process. Symposium Presented at the Annual Meeting of the Southwest Educational Research Association, Houston, Texas,m 1986.

Ward, Annie W., Blackman, Margaret E., Hall, Bruce W. and Mazur, Joseph L. Guide for School Testing Programs. National Council on Measurement in Education appr. 1974.

Wilkens, Susan A. The Role and Function of Testing Units in Large School Districts. Unpublished Paper of the Fairfax County Public Schools, Fairfax, Virginia, 1981.