DOCUMENT RESUME

ED 241 587 TM 840 135

AUTHOR Frechtling, Joy A.; Myerberg, N. James
TITLE Reporting Test Scores to Different Audiences.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO ERIC-TM-85
PUB DATE Dec 83
CONTRACT 400-83-0015
NOTE 77p.; Some tables contain small print.
AVAILABLE FROM ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, NJ 08541 ($7.00).
PUB TYPE Guides - Non-Classroom Use (055) -- Information Analyses - ERIC Information Analysis Products (071)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Annual Reports; Elementary Secondary Education; Evaluation Utilization; Graphs; Guidelines; *Information Dissemination; Parent School Relationship; *Scores; Teachers; *Testing; *Test Interpretation; *Test Results

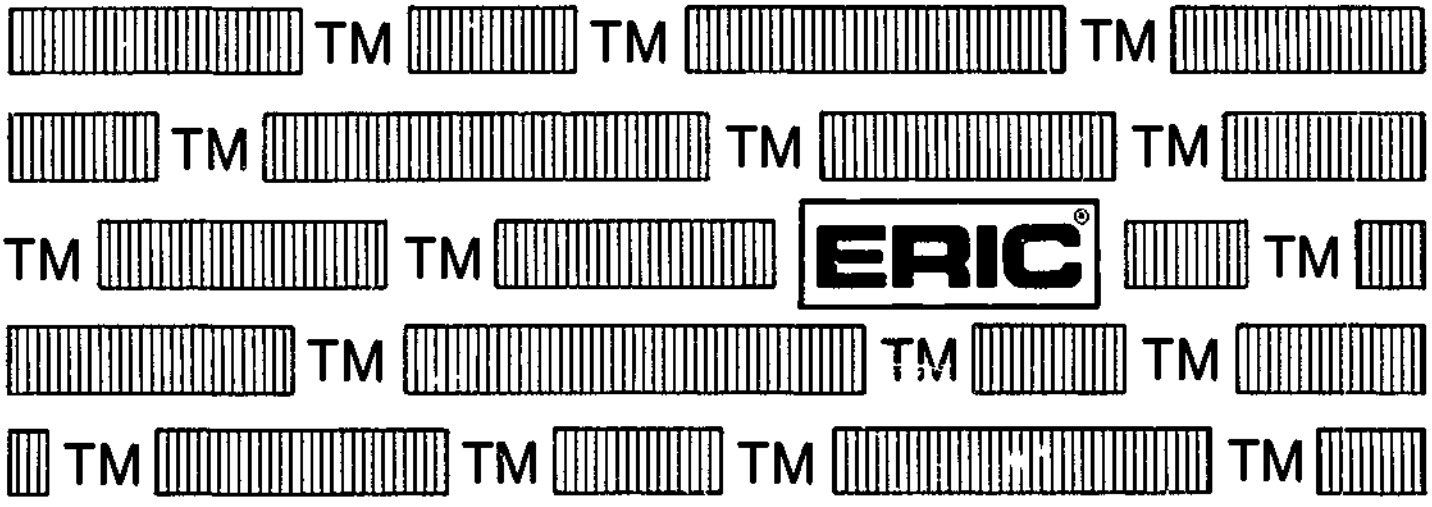
ABSTRACT

The purpose of this document is to address issues related to the release of test scores to a variety of audiences: parents, school board members, school staff, the news media, and the general public. Guidelines or recommendations for reporting test data are provided. The recommendations are based both on experiences in reporting test results and an informal review of a sample of test reports from school districts across the nation (see Appendix A). Annual reports on testing programs should include (1) descriptive information of the testing program, test content, and test scores; (2) test results for districts, as well as for individual schools; and (3) cautions concerning how the data should and should not be interpreted. Reports to parents will include the same information, but focused on an individual student. Reports to staff will focus on a class or a school. Suggestions for using test data for comparing schools, determining weak and strong areas, and determining if a school did as well as it should have are presented. Commonly used test terms, testing textbooks that include discussions of testing terms, and reports of test results cited in "Research and Evaluation Studies from Large School Districts 1982" are included in the appendices. (PN)

REPORTING TEST SCORES TO DIFFERENT AUDIENCES

ERIC/TM REPORT 85

by
Joy A. Frechtling
N. James Myerberg



ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08541

ERIC/TM Report 85

REPORTING TEST SCORES TO DIFFERENT AUDIENCES

by

Joy A. Frechtling

N. James Myerberg

Montgomery County Public Schools

Rockville, Maryland

December 1983

ERIC Clearinghouse on Tests, Measurement, and Evaluation

Educational Testing Service, Princeton, NJ 08541-0001

The material in this publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to qualified professionals for critical review and determination of professional competence. This publication has met such standards. Points of view or opinions, however, do not necessarily represent the official view or opinions of either these reviewers or the National Institute of Education.

ERIC Clearinghouse on Tests,
Measurement, and Evaluation
Educational Testing Service
Princeton, NJ 08541-0001

Table of Contents

List of Exhibits

Chapter 1

INTRODUCTION

Ten years ago, the practice of releasing test scores to the public was not generally accepted. A study by the Educational Research Service, conducted during the 1973-74 school year, showed that only 52 percent of the school systems enrolling 12,000 or more pupils released standardized test scores to the press (ERS, 1974). At about this same time, a "how to" publication by the National School Public Relations Association (NPRA, 1976) introduced a chapter on one state's experience in releasing test scores with the following admission:

> "Quite candidly, those associated with the Maryland Department of Education in 1974 approached the first time release of test results in panic."

The situation in 1983 is quite different. The release of test scores to the press and the general public is a common practice. Test scores are considered a statistic in the public domain similar to population estimates and tax rates. In some cases, public libraries even include school districts' annual test reports among their general reference materials.

The issue today is not whether or not to release test scores, but rather what to release and how to release it. Further, it has been increasingly acknowledged that since the audience for test scores has different faces with different backgrounds or iterests, the content and format of reporting may also need to be varied.

The purpose of this document is to address issues in the release of test scores to a variety of audiences: parents, school board members, school staff, the news media, and the general public. In the chapters which follow we will discuss the kinds of information that such reports might include and suggest some strategies for presenting them.

Before turning to the issue of how to report test scores, it is important to consider the question of exactly what one is trying to communicate. What information is the school district trying to get across? On the surface, this question has a simple answer: the purpose of reporting test scores is to tell an audience how well students did on some type of test. However, there is a second and equally critical purpose of reporting test scores: to provide the audience with an understanding of what test scores really mean and what they do not mean. This is a harder task for all involved.

In the chapters which follow, we will look more closely at the issues involved in reporting test scores: the kinds of information to be reported, the reasons for including each, and some ways in which the information might be presented.

Chapter 2

OVERVIEW OF THIS REPORT

In the next several chapters, we describe the kinds of information that a report on test scores might include. Although most districts actually have several different reports on testing (reports to the board, reports to parents, reports to school-based staff, etc.), we will begin with the annual test report, the report issued to the board of education and the public, as it is typically the one which is the most formal and complete. This is also the report that is most widely read and usually forms the basis for the major press coverage that test scores receive. In subsequent chapters, we will talk briefly about reports to other audiences: parents and school staff.

Our aim in presenting this information is to provide guidelines or recommendations for reporting test data rather than a set of prescriptions. Although our discussion will cover a wide range of areas, we recognize that not all are likely to be included in reports by individual school districts. Factors such as practical limits on the kinds of data which are readily available and the political sensitivity of the information may well affect what is included.

Our recommendations are based both on our experiences in reporting test results and an informal review of a sample of test reports from school districts across the nation (see Appendix A). Although we do not claim that these reports are either exemplary or representative of current practice, they provided us with valuable insight into how

different districts have approached the problem as well as some practical examples of how information is communicated. They also offer clear evidence that although some consistent themes emerge, there is no one way of doing things; both content and format differ considerably.

In the next several chapters, we discuss three areas of information which we feel should be included in some way in an annual report on testing. These are:

1) Descriptive Information

2) Test Results

3) Interpretive Cautions

Where possible, examples which we feel are useful from annual test reports by school district have been included as illustrations.

Chapter 3

GENERAL REPORTS: DESCRIPTIVE INFORMATION

Three kinds of descriptive data should be included in a report: a description of the testing program, a description of what the tests measure, and a description of the test scores. Although these sound like very basic and simple elements, review of existing reports indicates that they are not always included.

DESCRIPTION OF THE TESTING PROGRAM

A brief description of the testing program includes the names of the tests used, how they were developed and/or normed, when the tests were administered, and the grades in which they were administered. The test name should include the form and/or levels used, to facilitate comparisons with other test results. Information on norming should include when the test was normed and whether separate norms are provided for special subgroups, e.g., large cities or suburban districts. Provision of administration dates is also useful, both to help in this comparison and to indicate whether testing occurred at the beginning or end of the school year. Exhibit 1 shows how the San Diego City Schools describes its testing program in its report for the 1981-82 school year. Included are data on the tests used, grades tested, dates administered, and content covered.

Additional information which may be offered includes data on exemption criteria and percentage of students tested. These data can be very important. The same test score may well be interpreted quite

EXHIBIT 1
DESCRIPTION OF TESTING PROGRAM
(SAN DIEGO CITY SCHOOLS)

TESTS ADMINISTERED AND DATES

During the 1981-82 school year, state and nationally standardized tests were administered districtwide to approximately 50,000 San Diego students in Grades 3, 5, 6, 7, 11, and 12, to obtain data for two testing programs. The programs are the California Assessment Program and the Districtwide Testing Program. The types of tests and the testing periods for these two testing programs were as follows:

California Assessment Program

Survey of Basic Skills: Grade 3 administered in late April and early May 1982, covering content areas of Reading, Written Language, and Mathematics.

Survey of Basic Skills: Grade 6 administered in April 1982, covering content areas of Reading, Written Language, and Mathematics.

Survey of Basic Skills: Grade 12 administered in December 1981, covering areas of Reading, Written Expression, Spelling, and Mathematics.

The California Assessment Program test at Grade 12 was identical to the test used the previous six years. The new third grade test was administered for the third time this spring. Previously, Grade 3 pupils were tested only in the content area of Reading. At Grade 6, a new test was administered this spring for the first time.

Districtwide Testing Program

Comprehensive Tests of Basic Skills, Level G, Form U, administered to Grade 5 students in April 1982, covering curriculum areas of Reading, Language, and Mathematics (reported October 12, 1982, Report 330).

Comprehensive Tests of Basic Skills, Level H, Form U, administered to Grade 7 students in April 1982, covering curriculum areas of Reading, Language, and Mathematics (reported October 12, 1982, Report 330).

Comprehensive Tests of Basic Skills, Level 4, Form S, administered to Grade 11 students in November 1981, covering curriculum areas of Reading, Language, and Mathematics (reported this spring, Report 305).

The tests administered for Districtwide Testing Programs at the elementary and junior high school levels were changed to different grade levels in recent years to reduce the amount of instructional time consumed by testing. Also, the district program changed from CTBS, Form S to CTBS, Form U this spring. More details may be found in Report 330.

This exhibit illustrates one approach to describing a testing program. It includes information on the tests used, the grades testing, dates of test administration, and the content areas covered.

differently where 40 percent of the students have been tested as opposed to 95 percent. It may be especially important to include information on who is exempted and the percentage of students actually tested where a district or school contains significant numbers of special education students or students of limited English proficiency. Exhibit 2 shows one format for reporting exemption data which is used by the Dallas Independent School District. Data are presented by both race and exemption criteria.

DESCRIPTION OF TEST CONTENT

This section should include descriptions of the specific skills measured by each subtest and how the skills are measured (i.e., item format). A discussion of the skills that are measured is needed because subtest names frequently reflect the favorite jargon of a particular test publisher and convey little meaning to someone not thoroughly familiar with the specific test battery. Sometimes the subtest name uses highly technical terms and requires formal understanding of an area, such as the subtest name, Structural Analysis (used on the California Achievement Tests). Other times, the name may cover so many skills that the specific ones being measured need to be stated. An example of this is Mathematics Concepts (also used on the California Achievement Tests).

Exhibit 3 shows how the Washington, D.C., Public Schools describes what is included in the Comprehensive Tests of Basic Skills in their 1982 report on test scores. This report provides, in addition to a description of the test and subtest content, information on the number of items included in each subtest at each grade. An alternative approach

EXHIBIT 2
PRESENTATION OF EXEMPTION DATA
(DALLAS INDEPENDENT SCHOOL DISTRICT)

Summary of Exemptions from Components of System-wide Testing Program

Grade	All Students					White					Black					Hispanic				
	SE	LEP	SD	TOT	%	SE	LEP	SD	TOT	%	SE	LEP	SD	TOT	%	SE	LEP	SD	TOT	%
NORM REFERENCED TESTS[a]																				
K	95	350	3	448	4.9	21	6	0	27	1.3	53	3	2	58	1.3	21	329	1	351	14.4
1	261	1121	13	1395	12.1	110	6	1	117	4.1	107	2	5	114	2.2	43	1105	7	1155	35.4
2	402	784	4	1190	11.5	155	1	1	157	5.9	179	5	0	184	3.8	62	773	3	838	31.5
3	565	584	3	1152	11.2	216	3	3	222	8.5	232	1	0	233	4.8	117	573	0	690	26.3
4	657	198	9	864	8.7	222	0	2	224	8.6	302	1	2	305	6.2	125	196	4	315	14.2
5	735	176	4	915	8.7	232	1	0	233	8.5	367	0	1	368	7.0	133	167	3	303	13.2
6	756	160	14	930	9.4	207	0	4	211	7.9	437	0	9	446	8.6	108	151	1	260	13.1
7	614	75	44	733	7.4	166	0	15	181	6.8	350	0	26	376	7.6	96	51	3	150	7.5
8	589	50	11	650	6.6	132	1	0	133	5.1	367	0	10	377	7.1	88	35	1	124	6.9
9	553	212	0	765	7.3	151	1	0	152	5.2	336	2	0	338	6.1	65	109	0	174	9.9
10	515	41	0	556	5.9	162	0	0	162	5.9	294	1	0	295	5.8	54	33	0	87	6.5
11	388	16	0	404	5.2	138	0	0	138	5.4	204	0	0	204	4.9	42	15	0	57	5.8
ASSESSMENT OF BASELINE CURRICULUM (ABC)																				
1	261	283	4	548	4.7	110	3	1	114	4.0	107	2	2	111	2.1	43	273	1	317	9.7
2	401	168	3	572	5.6	153	0	1	154	5.8	100	3	0	103	3.8	62	162	2	226	8.5
3	564	167	3	734	7.1	214	0	3	217	8.3	233	0	0	233	4.8	117	161	0	278	10.6
TEXAS ASSESSMENT OF BASIC SKILLS (TABS)																				
3	557	0	0	557	5.4	215	0	0	215	8.2	228	0	0	228	4.7	114	0	0	114	4.3
5	724	0	0	724	6.9	228	0	0	228	8.3	362	0	0	362	6.9	131	0	0	131	5.7
9	549	0	0	549	5.2	148	0	0	140	5.1	336	0	0	336	6.1	64	0	0	64	3.6
10	500	0	0	500	5.3	159	0	0	159	5.8	282	0	0	282	5.5	54	0	0	54	4.1
11	382	0	0	382	4.9	137	0	0	137	5.3	200	0	0	200	4.8	41	0	0	41	4.2

[a] Includes MAT, ITBS, and TAP; not CTBS.

Note. SE = Special education exemption; LEP = limited English proficiency exemption; SD = skill deficiency exemption; TOT = total number of exemptions; % = percent of enrollment exempted.

This exhibit shows one method of reporting the number of students exempted from testing. For each of the tests administered, data are presented on the numbers of students exempted by exemption category as well as the racial/ethnic and grade-level characteristics of the students exempted.

EXHIBIT 3
TEST CONTENT DESCRIPTION
(DISTRICT OF COLUMBIA PUBLIC SCHOOLS)

Total Reading scores are obtained by combining the Vocabulary and Comprension scores. The Reading Vocabulary subtest measures student skill in determining word meaning from the context in which a word appears in a phrase. Reading Comprehension items require the student to read passages, letters, poems and articles, and then to answer questions requiring literal recall, identification of main idea, critical comprehsion, ability to draw conclusions and other reading skills.

Total Mathematics scores are obtained by comoining the Computation, Concepts and Application scores. The Mathematics Computation subtest contains items requiring addition, subtraction, multiplication and division of whole numbers, fractions, decimals and algebraic expressions. Mathematics Concepts measures the student's ability to convert concepts expressed in one numerical, verbal or graphic form to another form and to comprehend numerical concepts and their interrelationships. Finally, Mathematics Application items measure the ability to carry out problem solving operations.

Total Language scores are obtained by combining the Language Mechanics, Language Expression and Spelling scores. The Language Mechanics subtest measures student skill in capitalization and punctuation. The Language Expression subtest measures correctness and effectiveness of language usage, diction, economy and clarity of expression, and skill in organization. The Spelling test measures the ability to recognize spelling errors.

The Reference Skills test assesses knowledge of the uses of a library, parts of books and standard reference works. The Science items are related to the various content areas of the physical and life science.

Exhibit 3, continued

Tests	Subtests	Number of Items by Grades 3	6	9	11
Reading	Test 1 - Reading Vocabulary	40	40	40	40
	Test 2 - Reading Comprehension	45	45	45	45
Spelling	Test 3 - Spelling	50	50	30	30
Language	Test 4 - Language Mechanics	20	20	20	20
	Test 5 - Language Expression	35	35	35	35
Mathematics	Test 6 - Mathematics Computation	48	48	48	48
	Test 7 - Mathematics Concepts and Applications*	50	50	50	50
Reference Skills	Test 8 - Reference Skills	20	20	20	20
Science	Test 9 - Science	36	36	41	40
Social Studies	Test 10 - Social Studies	37	37	40	30

* Separate scores are reported for Concepts and for Applications.

This exhibit provides an example of how one district presents detailed information on exactly what its testing program assesses. The text provides a brief description of what the subtests measure and the table shows how much attention is devoted to each of the general areas.

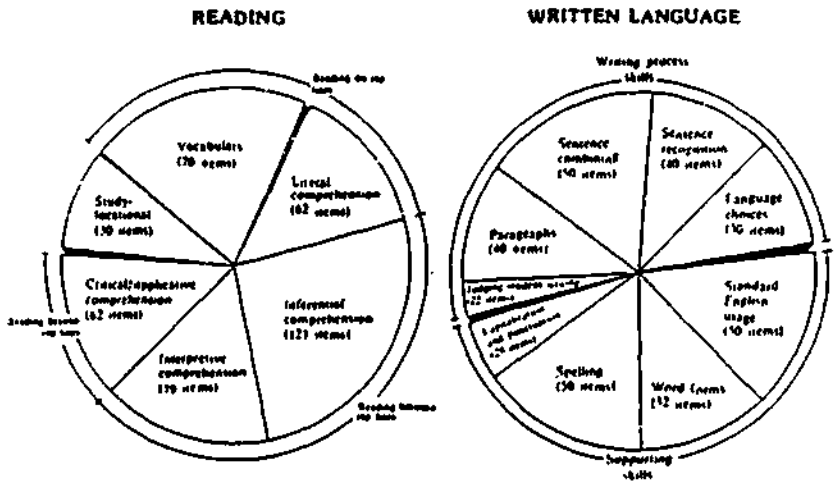
taken by the San Diego City Schools is presented in Exhibit 4. The use of pie graphs to display this information is somewhat unusual, but clearly communicative.

DESCRIPTION OF TEST SCORES

The final area is that of description of test scores. This is where the metric being used to report test data is presented and should be defined. The contents of this type of discussion will clearly vary depending upon the actual scores being used, e.g., percentile ranks, grade equivalents, etc., and the kinds of test being considered, e.g., criterion-referenced vs. norm-referenced tests. The critical factor is the presentation of some definition after the metric is introduced. Exhibit 5 presents an excerpt from the 1981-82 test report of the Houston Independent School District, in which definitions are provided for two different metrics used in reporting their test scores: grade equivalent scores and percent mastering each objective. These descriptions not only provide a clearly understood definition for each of the terms but also suggest possible pitfalls in their interpretation. This important area will be discussed in greater detail in a later chapter. Appendix B presents definitions for some commonly used test terms with cautions concerning their usage. Several elementary testing textbooks also have considerable discussion of these terms. See Appendix C for a list of these books.

EXHIBIT 4

PIE GRAPH SHOWING CONTENT DISTRIBUTION
(SAN DIEGO CITY SCHOOLS)



This exhibit illustrates an alternative way of describing what is assessed by a testing program using pie charts rather than text and tables.

EXHIBIT 5
EXPLANATION OF SCORES USED IN A TEST REPORT
(HOUSTON INDEPENDENT SCHOOL DISTRICT)

How are the results of the tests reported?

For the ITBS (Grades 1-6), all scores are reported in terms of grade equivalents. The grade equivalent (GE) on any test/subtest represents the grade level at which the "typical" pupil made this score. The first digit of a GE score represents the grade level and the second digit represents the month within the grade in which the typical pupil answered a particular number of questions correctly. For example, if a student earns a GE of 6.3, his number of right answers was the same as that typically made by pupils in the sixth grade at the end of the third month. The GE should be regarded as an estimate of where the pupil is along a developmental continuum, not where he/she should be placed in the grade organization of the school.

The TABS results are reported in terms of the percent of students mastering each objective. For the reading and mathematics subtests, a student demonstrates mastery of an objective by correctly answering at least three out of four test items under the objective. The writing subtest requires students to write a paragraph on a selected topic in addition to answering a series of multiple-choice questions. The writing sample makes possible an assessment of the student's handwriting as well as a rating of the student's organization and appropriateness of response. The latter two components of the writing subtest are combined to produce a raw score ranging from zero to four. On this scale, zero is poor and four is excellent. To demonstrate mastery, students are required to receive a writing sample raw score of at least two.

TABS scores for Vanguard schools are not reported separately from the entire campus, therefore no separate TABS data are included for Vanguard program.

This exhibit illustrates one way of explaining the types of test scores used. Of particular note is the inclusion of cautions which must be kept in mind in interpreting the data.

Chapter 4

GENERAL REPORTS: TEST RESULTS

Annual test reports generally include two types of data: overall district results and results for individual schools. These are usually presented in a very similar fashion, using the same descriptors and addressing the same basic questions.

DISTRICT RESULTS

Annual reports on districtwide results commonly present two types of information: information on how well the typical or average student performs, and information on how performance differs among students. In addition, annual test results may be supplemented by historical data which assist in the interpretation of performance in any single year.

The particular metric method used for displaying average performance will vary depending on the type of test. In reporting data on norm-referenced standardized tests, average scores, reported in terms of stanines, percentile ranks, or grade equivalents, are generally presented. Sometimes districts also report the percentage of students scoring above some reference point, typically the national mean or median. In reporting scores on criterion-referenced tests, the results are usually presented in terms of percentage passing. While tables are frequently used to present these data, graphical displays are especially useful.

Information on how performance differs among students can be communicated by presenting a frequency distribution. One way to

accomplish this would be to report the percentage of students falling into each quarter of the national norms. This is the way in which the Albuquerque Public Schools presents such information (Exhibit 6). Another approach is to present the data using stanines which show the spread of scores in a little more detail than national quarters. Exhibit 7 shows how information on spread of scores was presented by the Montgomery County Public School system in their 1981-82 Annual Test Report. This exhibit not only provides information for the county but also includes comparative data from the national norm sample.

Another way to look at performance differences among students is to present test scores by socioeconomic status (SES), by the major racial/ethnic groups, and/or by gender. Although it is recognized that reporting such information can be politically sensitive, these data can be useful in identifying areas where special efforts may be needed. The formats for reporting described in the previous paragraph are equally applicable here. We would like to stress, however, that reporting score distributions may be especially important. Average scores may give the impression that students from different groups perform very differently. Although this may be true on the average, it is also important to show that most groups have some members with high scores and some with low scores, no matter how high or low their average scores are.

One caution in grouping students by SES must be mentioned here. SES information can be useful in helping audiences to understand test results, since standardized test scores have repeatedly been shown to be highly related to SES variables such as parental income, parental education, and parental occupation. However, while SES data provide a

EXHIBIT 6

REPORT OF STUDENTS SCORING IN
EACH NATIONAL QUARTER

(ALBUQUERQUE PUBLIC SCHOOLS)

CTBS FORM U

NUMBER AND PERCENT OF STUDENTS IN EACH PERCENTILE RANGE

SPRING, 1982

APS EIGHTH GRADE

	COMPREHENSIVE TESTS OF BASIC SKILLS													
	READ VOCAB	READ COMP	READ TOTAL	LANG MECH	LANG EXPR	LANG TOTAL	MATH COMP	MATH CONC & APPL	MATH TOTAL	BATT TOTAL	SPELL	REF SKILLS	SCIENCE	SOCIAL STU
76 TH THROUGH 99 TH PERCENTILE ▶	1544	1305	1405	1645	1743	1635	1713	1697	1726	1516	1318	1314	1330	1507
	30%	25%	28%	32%	39%	32%	33%	33%	34%	30%	26%	26%	26%	30%
							75TH PERCENTILE							
51 ST THROUGH 75 TH PERCENTILE ▶	1288	1556	1403	1214	1543	1335	1251	1470	1279	1323	1118	1248	1302	1399
	25%	30%	27%	24%	28%	26%	25%	29%	25%	26%	22%	25%	25%	27%
							50TH PERCENTILE							
26 TH THROUGH 50 TH PERCENTILE ▶	1158	1163	1300	1356	995	1294	1194	950	1234	1280	1391	1683	1590	1514
	23%	23%	25%	26%	18%	26%	23%	19%	24%	26%	27%	33%	31%	30%
							25TH PERCENTILE							
1 ST THROUGH 25 TH PERCENTILE ▶	1154	1119	1028	931	847	831	972	1004	873	917	1317	840	894	685
	22%	22%	20%	18%	16%	16%	19%	20%	17%	18%	25%	17%	18%	13

N = 5144

TABLE 2

CTBS PERCENTILE RANGE ANALYSIS

The above information can be used in two ways:

1. To determine how many students fall into the high, above average, below average, and low achievement percentile range groups.
2. To determine what percentage of the students fall into each of the four percentile range groups.

The actual number of students will be of more interest to persons responsible for providing special services or materials to a particular type of student.

The percentage of students falling in each percentile group in the national norming is always 25%. To the extent that your school percentages are above or below 25% in each category reflects whether you have a greater or lesser percentage of students in that group than the national distribution.

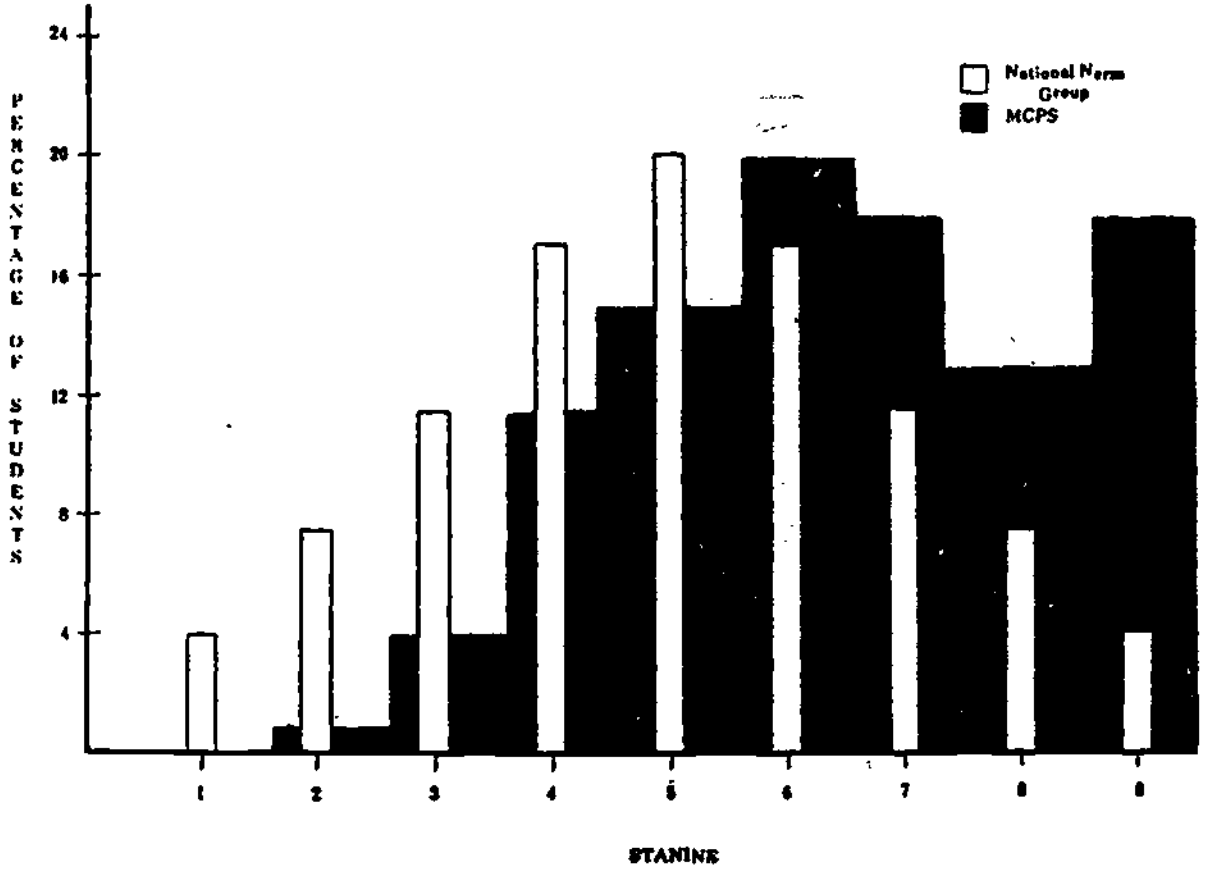
APS DISTRICT ——— Grade 8

This exhibit shows how data on the number and percent of students scoring in each quarter of the national norm group can be used to report the distribution of test scores.

EXHIBIT 7

GRAPHIC PRESENTATION OF DISTRICT
AND NATIONAL STANINE DISTRIBUTION
(MONTGOMERY COUNTY PUBLIC SCHOOLS)

CALIFORNIA ACHIEVEMENT TESTS, FALL 1981
DISTRIBUTION OF STANINE SCORES ON
THE TOTAL BATTERY FOR ALL GRADES TESTED



This exhibit shows an alternative way to report score distributions using a graphic, as opposed to a tabular, display. Using overlays, this exhibit also shows how data on local score distributions can be compared to those in the national norm sample.

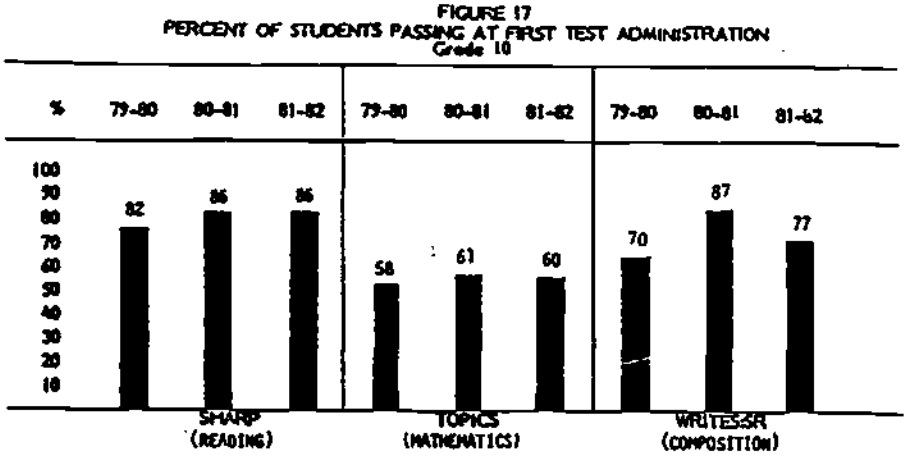
partial explanation for some antecendents of low test performance, such data must not be used to justify continued lack of academic success. In other words, such data should not be used to explain away the problem of low test performance nor relieve the school of the responsibility for trying to increase learning.

Our discussion so far has focused on reporting test scores for a single year. It can be useful to put such annual test results in a historical perspective to judge whether achievement is improving or declining. The historical data can be presented in one of two ways--cross-sectionally or longitudinally. Cross-sectional data show the results for each grade tested each year. All students tested in each grade are included. These results simply show if the scores for each grade in a given year were higher or lower. Exhibits 8 and 9 show two alternative ways of presenting such data: bar graphs used by the Los Angeles Unified School District (1981-82 test report) and line graphs used by the San Diego City Schools (1981-82 test report). The latter also compares city results to those of the state.

Since cross-sectional displays provide data for different students each year, any trends could be caused by changing student ability, not quality of instruction. To eliminate the possible changes in ability level, longitudinal data are needed. Longitudinal data show the trend of scores across two or more years for students tested in all years. This could show not only whether achievement was improving or declining but also provide some indication of the quality of instruction. However, to be able to do this, it is necessary to control for differences in the

EXHIBIT 8

REPORTING CROSS-SECTIONAL DATA USING
A BAR GRAPH (LOS ANGELES UNIFIED SCHOOL DISTRICT)



This exhibit shows one method of reporting historical, cross-sectional data using bar graphs to illustrate changes in performance over time.

EXHIBIT 9

REPORTING CROSS-SECTIONAL DATA USING
A LINE GRAPH (SAN DIEGO CITY SCHOOLS)

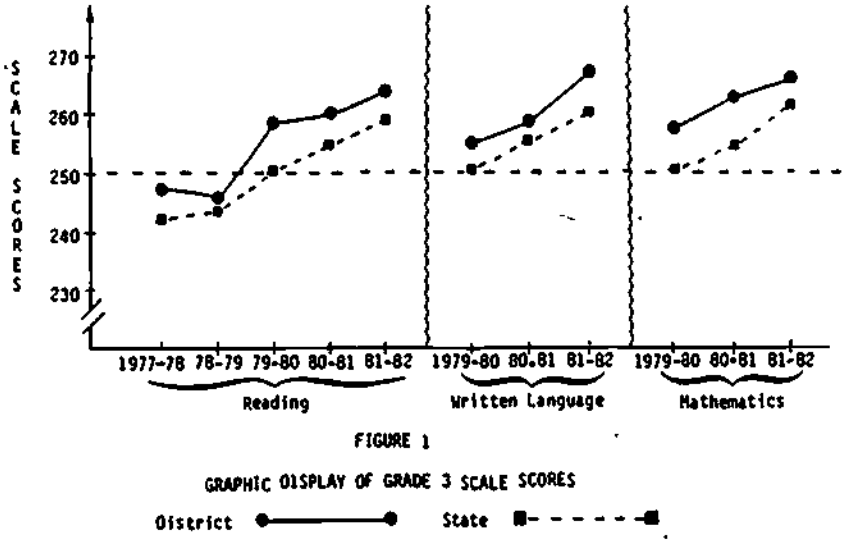


FIGURE 1

GRAPHIC DISPLAY OF GRADE 3 SCALE SCORES

District ●——● State ■- - - -■

This exhibit shows an alternative way to present historical cross-sectional data. In addition, in this exhibit, statewide data have been added to provide a reference group for the local data of interest.

tests at each grade level. This will be discussed in detail in a later chapter on how to use test data.

SCHOOL RESULTS

Average scores, percentage passing, and score distributions may be presented for each school in a district summary report in a fashion similar to that used in presenting districtwide data. It might be best to limit the distribution here to number and/or percentage in each national quarter to minimize the data that the reader has to deal with. A slightly different way to present the data is to show the scores for the student at each quartile in the school. Exhibit 10 shows how the Montgomery County Public Schools presented this information. School staffs may be sent more detailed frequency distributions, as well as data such as performance by objective, in a separate memo. Reports to school staffs will be discussed in more detail in a later chapter.

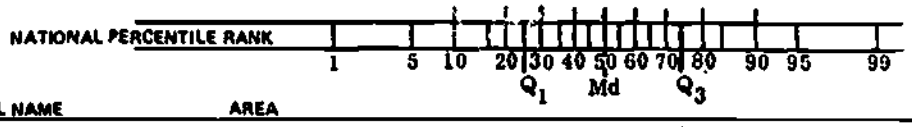
Results by race and sex for each school are also useful but should only be presented if the groups are large enough to provide good data. Since mean scores for small groups can be affected by a few extreme scores, reporting results by race or sex for small groups can lead to misinterpretation.

Historical data can also be useful for schools. In the case of schools, it is even more important to use longitudinal data than for the district because factors--such as SES and ability--that distort cross-sectional data have an even greater impact at the school level than at the district level. Once these factors are eliminated, it is much

EXHIBIT 10

USING BAR GRAPHS TO SHOW TEST SCORE SPREAD IN SCHOOLS
(MONTGOMERY COUNTY PUBLIC SCHOOLS)

NATIONAL PERCENTILE RANK FOR THE STUDENT SCORING AT EACH SCHOOL'S FIRST QUARTILE (Q1), MEDIAN, AND THIRD QUARTILE (Q3) – CALIFORNIA ACHIEVEMENT TESTS GRADE 5, TOTAL BATTERY, 1981-82



SCHOOL NAME	AREA			
ARCOLA ELEMENTARY	1	48	75	90
ASHBURTON ELEMENTARY	2	53	72	93
AYRLAWN ELEMENTARY	2	66	74	90
BANNOCKBURN ELEMENTARY	2	58	82	95
LUCY BARNSLEY ELEMENTARY	2	68	86	96
BELLS MILL ELEMENTARY	2	74	90	97
BELMONT ELEMENTARY	1	72	83	93
BEL PRE ELEMENTARY	1	55	75	92
BETHESDA ELEMENTARY	2	69	89	95
BEVERLY FARMS ELEMENTARY	2	68	91	97
BRADLEY ELEMENTARY	2	76	92	98
BROAD ACRES ELEMENTARY	1	33	43	63
BROOKHAVEN ELEMENTARY	2	53	74	91
BROOKMONT ELEMENTARY	2	69	87	97
BROOKVIEW ELEMENTARY	1	40	51	66

This exhibit shows how bar graphs can be used to provide information on the average score and distribution of scores for individual schools.

more likely that an increase or decrease in performance is related to the school program.

SES data for a school can be very helpful in evaluating its test results. Once again, it should be noted that SES factors should be used to help understand test results, not to justify low performance.

Chapter 5

GENERAL RESULTS: INTERPRETIVE CAUTIONS

One of the most difficult and frustrating aspects of reporting test results is to assure that the results are interpreted and used as accurately as possible. This is not an easy matter, as most people think they know what a test score means although very few people really do. An example of the confusion which too often surfaces can be illustrated using grade equivalent scores. What does a grade equivalent score of 7.2 mean? It means that a student is working on the level of a seventh-grader in his or her second month of school. Right? Wrong. But this interpretation sounds right and is far too commonly heard. In fact, some of the most dangerous and common misinterpretations occur where what sounds right or what makes good common sense is technically wrong. Unfortunately, these misinterpretations are very difficult to reverse.

In reporting test scores to the public it is, therefore, critical to provide cautions concerning how the data should and should not be interpreted. Exactly what these cautions are depends on the particular data being reported and the kinds of tests being used. Listed below are some suggestions for inclusion, gleaned from areas where misinterpretation has been noted to occur frequently. Considered are problems in

- comparing scores across test batteries
- comparing data across grade levels
- interpreting normative data

- comparing the performance of different groups of students
- interpreting small changes in test performance

Appendix B provides information on cautions to be observed in using various types of test scores. In this section, we present some additional cautions which should should be kept in mind when interpreting data.

There are problems in comparing scores across test batteries. People frequently want to compare scores across school districts where districts do not use the same tests. Such comparisons are based on the mistaken belief that most tests measure the same thing, achievement, and that a test called reading comprehension on one battery is approximately equivalent to a test called reading comprehension on another battery. This can lead to some erroneous conclusions. There are several reasons for this caution.

First, in norm-referenced tests (NRT), norms for each test are based on a different group of students who may themselves differ in ability. Although test developers attempt to obtain a nationally representative sample for their norming groups, actually obtaining such a sample has become increasingly difficult as more and more districts have refused to participate in such endeavors. We simply do not know how the norming groups for different tests compare or whether certain tests set a higher standard of performance than others. For criterion-referenced tests (CRT), the standard setting methods for different tests may create problems analogous to those created by the development of norms for norm-referenced tests.

Second, achievement tests differ in their content, despite the fact that the names of tests or subtests may sound the same. Further, item formats may differ even where the same objective is being assessed. Depending on the type and extent of differences, the actual similarity of what is tested may, therefore, vary widely. Exhibit 11 describes how different item formats are used for tests of the same or similar names on the California Achievement Tests and the Iowa Tests of Basic Skills.

Caution must be used in comparing results for different grade levels, even where the same test battery is being used. The problems described above in comparing scores across different achievement tests are also found, although in slightly reduced form, in comparing scores across different levels of the same test. Again, the norms are based on different groups of students and we do not know for sure that the norm group for one grade level was in fact similar in critical areas to the norm group for another. While common sense suggests that in a "nationally representative group" cohort differences balance out, we simply cannot say with certainty that this is true. This issue becomes especially troublesome where trends in performance across grade levels are used to make some sort of judgment about the quality of instruction.

In attempting to make comparisons across grade levels, one must also be concerned with a second problem: the comparability of match between test content and curriculum content at the grade levels examined. Because tests are designed to reflect a consensus regarding what might be considered a national curriculum, they naturally do not reflect all local curricula equally well. If this match varies across grades, performance

EXHIBIT 11
COMPARISON OF ITEM FORMAT ON TWO
STANDARDIZED ACHIEVEMENT TESTS

Spelling (ITBS)/Spelling (CAT)--The ITBS asks the student to find an incorrectly spelled word in a list of words. The CAT asks the student to find an incorrectly spelled word in a sentence. Neither test asks the student to actually spell words and could not within the constraints of the opitical scan format employed.

Vocabulary (ITBS)/Reading Vocabulary (CAT)--The ITBS asks the student to find words that mean the same as a given word. The CAT contains some questions asking for the same meaning and some asking for the opposite meaning. It also has a few questions involving words with multimeanings. In these questions, s definition is provided and the student has to find the sentence in which the word is used with that definition.

This exhibit describes how two different achievement tests approach the measurement of the same skill, Spelling. This illustrates the point that one cannot assume that two subtests measure exactly the same skill simply because the tests used the same skill name.

differences unrelated to the quality of instruction are likely to be found.

This issue is very important in light of the not infrequent finding that test scores appear to decline as one progresses through the school years. Such a finding is typically interpreted as indicating that students do more poorly the longer they have been in school. An alternative hypothesis is that there is greater fidelity between test content and curriculum content in the early grades than the later grades. In the later grades, course content becomes more variable and a good match is harder to find.

Standardized test norms relate a student's scores to those of a norm group which took the test in the past when the test was standardized, not to the current group of students being tested. When people see test scores for a given year and percentile ranks showing how students performed relative to a national sample, there is a tendency to assume that the two groups took the test at approximately the same time. This is not the case even for the most recent edition of tests. There is generally one norming sample used for each edition of a test, and--depending on when a test was normed--that sample may have taken the test one to seven or more years ago.

Comparison of results for different groups of students can lead to incorrect, sometimes harmful, conclusions. In the discussion above, we have considered how some aspects of the tests themselves can influence the performance of students and thus complicate the interpretation of results. However, even when two groups of students exposed to similar programs take the same test with the same norms or passing standard, it

is necessary to consider factors other than the scores themselves in interpreting the findings and drawing conclusions about factors such as the quality of instruction. The major factors to consider relate to the socioeconomic status (SES) of students, indicated by variables such as income, parental education, and parental occupation. All of these have been shown to be highly related to standardized test performance, with higher SES students tending to show higher test performance (other things being equal).

Data on SES are not always available, either because the school district does not have access to the information or because the school district feels that the data are too difficult or sensitive to collect. Thus, it is not always possible to partial out an SES effect. If this is the case, it might nonetheless be useful to point out the importance of the relationship as a partial explanatory variable. This may be especially important where other factors are likely to be confounded with SES. An example of such confounding occurs where data are reported by racial/ethnic group or by school.

We could find no district that reports results by SES groups. However, several did include some SES information in their reports. One of the most thorough examples of this kind of reporting is shown in Exhibit 12. This is taken from the Dade County District and School Profiles, 1982-83.

Small test score differences should not be used to make educational decisions. All test scores contain measurement error. This can be caused by many things including ambiguous questions, how the student feels when he/she takes the test, lucky guesses, or distractions

EXHIBIT 12

REPORTING SCHOOL SES DATA AND STAFF CHARACTERISTICS (DADE COUNTY SCHOOLS)

This exhibit shows one way of including some data on socioeconomic status (SES) in a report on test scores. The amount of data presented on SES is quite limited, however, as only data on the percentage of students with free/reduced lunch are included.

occurring while the test is being administered. For these reasons, one must be very cautious in interpreting small differences as indicating meaningful differences in instructional quality or knowledge of skills. This is especially true for individual student results, since the error in scores for individual students tends to be much larger than that for group data.

The problem is, however, equally important where larger groups of students are concerned. Small differences in scores for large groups of students may appear important because statistical tests indicate that they are significant and, thus, unlikely to be caused simply by error. Although this is true, in interpreting such findings, one must also consider whether or not the difference is really meaningful, i.e., does a difference of one percentile point, although statistically significant if enough students are involved, merit major panic or euphoria on the part of a school system? Perhaps a good test of importance can be made by assessing how much money or how much change a school system would be willing to spend or make to cause so small a change to occur.

Chapter 6

REPORTS TO PARENTS

In the previous chapters, we discussed issues to consider in presenting an annual test report. In this chapter, we turn to a second audience: parents. Reports to parents are, in many respects, quite similar to reports to boards of education and the general public. Despite their focus on an individual student rather than a group, they still should provide a description of the program, test results, assistance, and cautions in interpretation. Typically, however, reports to parents are presented quite differently, under the assumption that parents, as laymen, must be given the information in a form which is both briefer and easier to comprehend. The question and answer format is popular (see Exhibit 13 taken from materials used by the Dallas Independent School District) as are brochures; slide/tape presentations are often used to provide an overview, and graphs and other pictorial displays are frequently found. The trick here is to make the description brief and easy to understand without, at the same time, appearing to insult the intelligence of the audience.

The most difficult and most important part of the report to parents is presenting the information on how their child performed. In reporting actual scores, it is important to choose a metric which is relatively easy to understand, and which can be readily defined. In reporting scores from norm-referenced tests, stanines are a popular choice because they appear on the surface to meet the criterion of ready

EXHIBIT 13

QUESTIONS AND ANSWER FORMAT FOR PROVIDING PARENTS WITH INFORMATION ABOUT A TEST PROGRAM (DALLAS INDEPENDENT SCHOOL DISTRICT)

TABS

TEXAS ASSESSMENT OF BASIC SKILLS - TABS
(STATE-MANDATED)

WHEN?	FEBRUARY EACH SCHOOL ESTABLISHES SPECIFIC DATES WITHIN THE GIVEN PERIOD
WHY?	REVIEW EDUCATIONAL NEEDS OF TEXAS PROMOTE PLANS FOR MEETING NEEDS EVALUATE ACHIEVEMENT
WHO?	ALL STUDENTS IN GRADES 3,5,9 AND OTHER STUDENTS IN GRADES 10, 11 AND 12
WHAT?	MINIMUM SKILLS MEASURE OF READING, WRITING, MATHEMATICS
SCORES?	•RESULTS REPORTED TO STUDENT, PARENT OR GUARDIAN, AND SCHOOL PERSONNEL •PROCESSED BY TEXAS EDUCATION AGENCY (AUSTIN). RETURNED MAY

This illustrates one method for providing parents with a description of a testing program. It is well suited for the purpose of communicating with parents because the question-and-answer format provides clear and quick answers to frequently asked questions.

comprehensibility. As long as no one actually asks for a definition, stanines may be a safe choice. However, many a test director has squirmed his/her way through several uncomfortable minutes after a PTA member has innocently asked, "What exactly are stanines?"

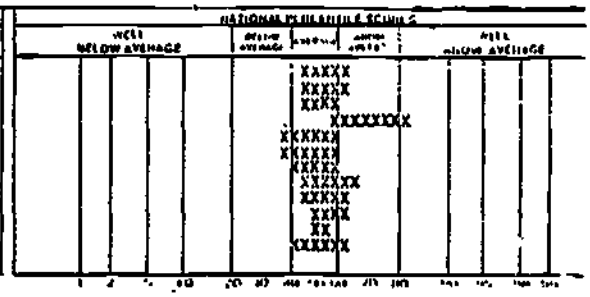
An alternative which may better serve communication are national percentile ranks. Since they have a range of 1 to 99, they fall on a scale which seems both familiar and easy to use. Although they may appear to convey greater precision than is justifiable, this fault is not unique to national percentile ranks. In fact, regardless of the method used for reporting individual scores, a relatively strong statement should be included regarding the error in test scores and their limitations.

Perhaps the single most critical thing in reporting to parents is conveying the message that test scores are far from perfect indicators of what a student has or has not learned. Materials accompanying such reports should, therefore, be quite clear about the multiplicity of factors that test scores may reflect. The fact that test scores typically contain a good deal of imprecision cannot be overstated. Unfortunately, the notion that an achievement test provides a precise measure of learning is all too widely held. A good way to get across the idea of test error on norm-referenced tests is to report scores using score bands as shown in Exhibit 14. This is part of the report of individual results used by the Pittsburgh Public Schools. This format for presentation reinforces the concept that a test score is not really a single point score, but an approximation. Such a display also helps to curtail concern over a change in performance of one or two points.

EXHIBIT 14

USE OF ERROR BANDS AND TEXT TO
REPORT INDIVIDUAL STUDENT RESULTS TO PARENTS
(PITTSBURGH PUBLIC SCHOOLS)

TEST	RS	NS	GE	NP
READING VOCABULARY	21	5	4.7	50
READING COMPREHENSION	24	5	4.7	50
TOTAL READING	45	5	4.7	50
SPELLING	15	6	6.5	73
LANGUAGE MECHANICS	15	5	4.6	49
LANGUAGE EXPRESSION	28	5	4.7	51
TOTAL LANGUAGE	43	5	4.7	50
MATH COMPUTATION	25	5	4.9	56
MATH CONCEPTS & APPL.	27	5	4.8	52
TOTAL MATH	52	5	4.8	53
TOTAL BATTERY	134	5	4.7	50
REFERENCE SKILLS	19	5	4.9	52



SUMMARY OF PUPIL'S SCORES
THIS STUDENT'S ACHIEVEMENT IN BASIC SKILLS MAY BEST BE SUMMARIZED BY LOOKING AT THE TOTAL SCORES. IT CAN BE SEEN THAT HER TOTAL SCORES ARE BETTER THAN APPROXIMATELY 50 PERCENT OF THE NATION'S 4TH GRADERS IN READING, 50 PERCENT IN LANGUAGE, 53 PERCENT IN MATHEMATICS, AND 50 PERCENT IN TOTAL BATTERY.

SHE HAS STRENGTHS IN CORRECTLY CAPITALIZING THE PRONOUN I, PROPER NOUNS, AND PROPER ADJECTIVES, IN SELECTING THE CORRECT PRONOUN AND ADJECTIVE FOR A SENTENCE, AND IN OBTAINING INFORMATION FROM THE INDEX OF A BOOK.

SHE HAS WEAKNESSES IN SPELLING ALL THE SOUNDS IN A WORD, IN CORRECTLY USING COMMAS AND QUOTATION MARKS, IN SOLVING COMPUTATION PROBLEMS IN DIVISION, AND IN UNDERSTANDING NUMBER SENTENCES.

SHE MAY NEED TO REVIEW OBTAINING INFORMATION FROM A DICTIONARY PAGE.

COMMENTS:

I have reviewed this report and have made additional comments where necessary.

Signed_____

DEAR PARENT:

This is a report of your child's test results in the basic skills. It shows you how well your child did on this year's tests. Your child is compared to other students in the same grade who took the tests throughout the nation.

The subjects tested are listed on the left side of the chart. The percentile scores are the percentages of students in the nation who scored below your child on each test. (See back of page for explanation of other scores.)

On the right side of the chart, the rows of Xs show how well your child did on the tests as compared to other students throughout the nation. Your child's national percentile scores are within the range indicated by the rows of Xs.

Also provided is further written explanation of your child's achievement as measured by these tests.

Please see your child's teacher [illegible]

This illustrates an effective way of providing parents with a report of individual student progress. Of special importance is the use of test score bands which readily illustrate the error that is part of test scores. This report format is reprinted by permission of the publisher, CTB/McGraw-Hill, Inc., 2500 Garden Road, Monterey, CA 93940.

The Pittsburgh report also shows how parents can be helped with a few lines of text highlighting individual strengths and weaknesses. Here the information on subtest performance has been supplemented by information on particular objectives in order to make the data more meaningful. This is useful as long as the test includes a sufficient number of items per objective and possible varying difficulty level of items and degree of objective coverage have been taken into account. Without these controls, such data on objectives may be misleading.

Finally, in reporting to parents, it is critical to keep the particular needs of this varied audience in mind. Presenting student results in simple English, free of jargon, may not be enough. More and more school districts are providing reports in languages other than English where substantial numbers of parents are likely to have limited English skills. An example from the Dade County Schools is shown in Exhibit 15. While it is debatable whether or not non-English alternatives should also be provided for reports such as the annual reports of districtwide results, it is far clearer that reporting in other languages is important where individual students are concerned.

It should be pointed out that not all school districts choose to send a written report on test scores to parents. As an alternative, scores are often conveyed verbally in some form of parent-teacher conference. This approach has the advantage of providing the opportunity for discussion between the parent and teacher and the chance for specific questions to be raised and addressed. Unfortunately, because not all teachers understand test scores equally well, it also sets the scene for some widespread miscommunication which may go uncorrected and undetected.

EXHIBIT 15

REPORTING RESULTS TO PARENTS IN TWO LANGUAGES (DADE COUNTY SCHOOLS)

SCHOOL

PARENT REPORT
STANFORD ACHIEVEMENT TESTS
DADE COUNTY SCHOOLS, 19 - 19

STUDENT PROFILE FOR

GRADE -

ID -

Dear Parent:

The Stanford Achievement Test I are administered to nearly all students in the Dade County Public Schools each year. These tests compare an individual's performance in the basic skills to the average performance of other pupils of the same grade throughout the nation.

Below is a report of your child's scores on the tests and an explanation of the results. If you have any further questions, please contact your child's school.

Estimados Padres:

Las pruebas de aprovechamiento Stanford Achievement Tests son administradas anualmente a casi todos los estudiantes de las escuelas del Condado de Dade. Estas pruebas comparan el aprovechamiento de conocimientos básicos en el individuo con el promedio de otros estudiantes en los mismos grados a través de la nación.

Debajo hay un reporte de los resultados de su hijo(a) en las pruebas y una explicación de los resultados. Si tiene alguna otra pregunta, por favor llame a la escuela de su hijo(a).

		LEVEL OF PERFORMANCE *Nivel de Aprovechamiento*				
	(7) PERCENTILE *Percentil*	(1) VERY LOW TO LOW *Muy bajo a bajo*	(2) BELOW AVERAGE *Por debajo del promedio*	(3) AVERAGE *Promedio*	(4) ABOVE AVERAGE *Por encima del promedio*	(5) HIGH TO SUPERIOR *Alto a superior*

(1) Requires remedial attention and increased instructional time at this skill

(2) Could benefit from remedial attention and more instructional time at this skill

(3) Should function satisfactorily within the regular grade level program

(4) Should progress more rapidly than the average pupil in programs requiring this skill

(5) Could benefit from advanced academic programs requiring this skill

(7) Percentile scores compare the student's performance with that of the national population. The average percentile score for each grade level in the nation is 50. For example, a percentile score of 46 for an eighth grader indicates that 46 percent of the pupils in the national eighth grade population made scores equal to or lower than the percentile score.

(1) *Requiere atención especial para reforzar este [illegible] adicional en esta asignatura*

(2) *Puede beneficiarse con ayuda especial adicional para [illegible] adicional en esta asignatura*

(3) *Debe funcionar satisfactoriamente dentro del programa regular*

(4) *Debe progresar más rápidamente en programas que requieren [illegible] que el estudiante promedio*

(5) *Podría beneficiarse de asignaturas en programas académicos [illegible] que requieren esta asignatura*

(7) *Los percentiles comparan el funcionamiento del estudiante con el [illegible] nacional. El percentil promedio para cada grado [illegible] promedio en este grado en la nación es 50. Por ejemplo, un [illegible] de 46 en octavo grado indica que el 46 por ciento de los alumnos en la nación [illegible] de octavo grado [illegible] una nota igual o más baja que esta [illegible]*

MHS-1137a7 Rev. [illegible]

This exhibit illustrates one approach to communicating test scores to parents with limited or no English skills. The critical information on student performance is printed in two languages, English and Spanish.

It may be that some combination of a formal, consistent, written communication and a personal conference with the teacher regarding the specifics of the classroom situation provides the safest and best way of reporting test score information to parents.

Chapter 7

REPORTS TO STAFF

The final audience to be considered here is school-based staff. This audience is composed of people serving several different functions: teachers, counselors, principals, and other specialists. Each has a slightly different use for test score data and each wants data presented in a slightly different form. While the approaches we have already discussed--the annual test report and the report to parents, the brochures and slide/tape presentations--partially meet these needs, they are not sufficient. Other data and other formats are better suited where staff use test data for program assessment and instructional decision making.

Typically, districts provide this additional information through school level reports or printouts which are intended primarily for use by staff of a particular school and are not commonly shared with other schools or the public. These are data displays rather than complete reports; they assume a fairly knowledgeable audience and frequently have little text or accompanying explanatory materials.

The exact contents and number of such reports, again, vary. One district sends out as may as twenty different reports on testing to schools annually. Others get by with far fewer. Information needs appear to be dictated not only by accepted conventions, but also by the specific concerns of a system in a given year. The list below provides an idea of the variety of kinds of reports that may be sent to schools:

Individual Student Reports. These reports are similar to those provided to parents. They list the total and subtest scores for each student. There are often several copies of these so that teachers and counselors can each have a copy.

Performance by Individual Classroom. These reports list the scores for students aggregated to the classroom level.

Frequency Distributions. These reports provide a detailed description of the spread of scores by including the number of students achieving each possible score.

List of High and Low Performing Students. These reports supplement the frequency distributions by indicating which students have exceptional scores. This report might be of use as part of the selection procedure for special programs or for grouping students for instruction.

Performance by Objective. These reports show how well some aspects of the curriculum are being mastered. When using these results, the number of items for assessing each objective, the difficulty of those items, and the extent to which the items measure the stated objective must be considered.

Performance Across Years. These reports provide historical summaries, either cross-sectional or longitudinal, of achievement over time. The information they provide can be useful for determining changes in school and student performance.

Performance by Feeder School. These reports show how well students from different feeder schools performed and provide information for the receiving school to use in planning the instructional program for these students.

Performance by Special Program. Part of the evaluation of special programs (e.g., Chapter 1, ESL, etc.) is to look at the test results of students in those programs.

Frequently, these data are presented or at least reviewed in a workshop-type setting using a variety of materials. This approach seems favored over attempting to include all information in a self-contained document, as is the case with reports to boards of education or parents. These workshops serve a dual purpose. They permit testing personnel both to communicate the information and to assure that the information is

being interpreted correctly. In addition, they allow school and staff to pose questions to the testers, which can lead to new analyses and/or better use of the information. Ideally, all audiences should have the opportunity to discuss test scores and receive assistance in their interpretation. However, it is especially critical that this occur with school staff. It is school staff that actually make decisions regarding individual students based on test data and it is at the school level that the impact of misinterpretation is the greatest. In large school districts, meeting with each school each year to go over school-level data may be overly ambitious, and some sort of staggered schedule may be more practical. The critical thing is that school staff receive sufficient opportunity for discussion and that reports to this audience evoke interaction as well as comprehension.

Chapter 8

SUGGESTIONS FOR USING TEST DATA

The previous chapters presented the elements that might be included in a report of test results and suggested some alternative strategies for presenting information to different audiences. this chapter will offer some suggestions about how one might go about answering questions regarding test scores that are frequently asked by all of these audiences. Here we are talking primarily about how the types of data described earlier might be used to respond to some of the more common questions posed. The emphasis here is on interpreting the data rather than simply reporting them. Three commonly asked questions are listed below.

1. How do a school's test scores compare with those of other schools?
2. In what areas does the school need to improve?
3. Did the students in the school do as well as they should have?

COMPARING SCHOOLS

Although one might wish it were not the case, one of the most popular uses of the data described earlier is to draw comparisons among schools, in the hope of making some assessment of school quality. Although using test data for this purpose is fraught with interpretive problems, there are clearly more or less acceptable ways of approaching this task. Too frequently, comparisons of schools are made simply by looking at test scores and determining which are higher or lower. In the

extreme, this leads to a ranking from top to bottom with the high-scoring schools considered "good" and lower-scoring schools considered "bad."

This approach can be extremely misleading because it totally ignores other factors affecting test performance and attributes all variance to the school. Unfortunately, we do not know of any totally satisfactory way of using test data to determine which schools are effective and which schools are not. However, suggested below are approaches which clearly improve on the simple ranking method described above.

Combining Test Scores and SES Data. Since standardized test scores are highly related to SES variables, it is likely that a school or group of students with low SES will also have low test scores. Thus, this relationship has to be accounted for so that schools with low or declining SES are not automatically labeled as instructionally inadequate. To avoid this kind of labeling, schools can be grouped according to SES. The test scores of schools within each group can then be compared to see how well each school is performing. An alternative approach is to use regression analysis to combine SES variables and produce "predicted" scores for each school and then to see which schools perform substantially above or below this prediction. The critical point here is that comparisons are made only among schools with students from similar backgrounds. Again, however, one must repeat the caution regarding the possible misuse of analyses which incorporate SES. SES data can be used to help understand test results, but they should never be used to provide a rationalization for tolerating low performance.

Longitudinal Analysis. Given the strong relationship between test scores and SES factors and the potential danger of using low SES to

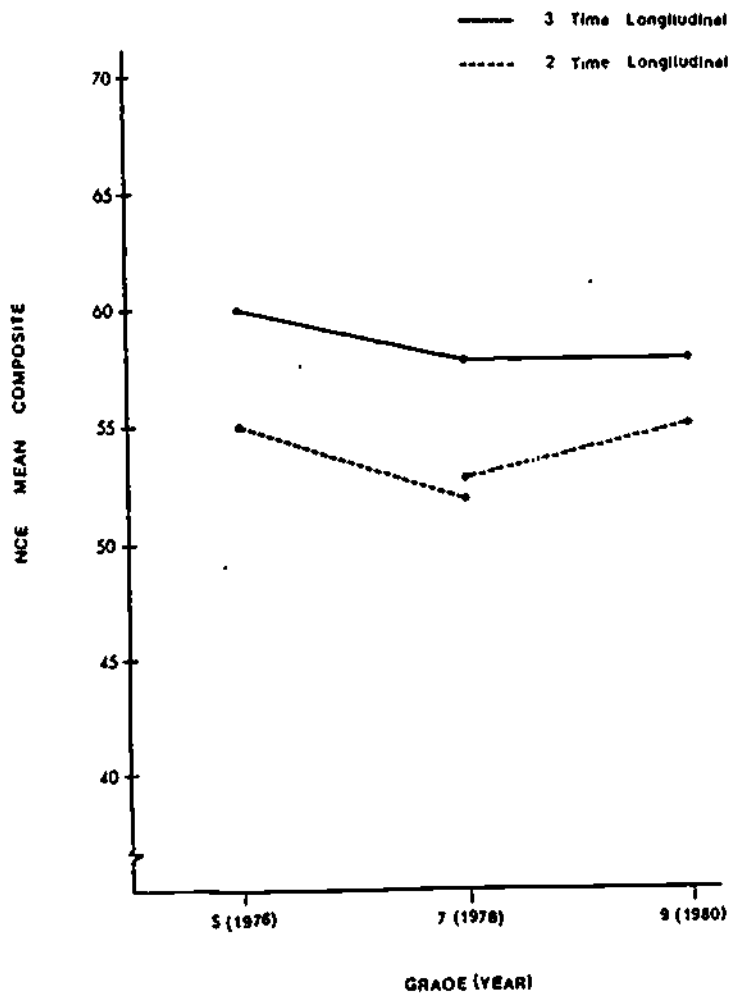
justify low performance, it may be better to deal with the qualitative issue in another way. Longitudinal analysis, introduced earlier, can be used to overcome the SES/test score relationship by using the same students at all data points and by using score trends instead of absolute values. However, one must keep in mind the possible problems posed by differences in tests and test norms discussed previously. To account for these differences, a baseline must be established. For example, the baseline could be established from the results of all students in a district who were tested in the same school in both Grade 2 and Grade 3. The trends of such students in each school could then be compared with this district baseline.

Although the longitudinal analysis described above provides a straightforward, fairly easy-to-understand way to use test scores to help make judgments about school programs, it does not make it possible to make the same judgment about the entire district, since it could be difficult to develop baseline data from a larger group with the same curriculum. About the best that can be done at the district level is to establish the baseline from the trends from one academic year and then compare the trends for all of the following years with that baseline.

Since longitudinal analysis involves looking at score trends, it provides an excellent opportunity to present the results graphically. Exhibit 16 shows the presentation of some longitudinal data in a report from the Montgomery County Public Schools.

EXHIBIT 16

GRAPHIC DISPLAY OF LONGITUDINAL RESULTS
(MONTGOMERY COUNTY PUBLIC SCHOOLS)



This exhibit illustrates how longitudinal data (data on the same students tested at two time points) can be used to display trends in test performance over time. This is proposed as an alternative way of presenting historical data.

DETERMINING WEAK AND STRONG AREAS

People frequently wish to know how well a school is performing in each academic area and what specific strengths and weaknesses exist. Suggested here is a way of determining strengths and weaknesses by comparing performance in each subject area to performance in all other subject areas. This method assumes that all subtests are part of the same test battery and no cross-battery comparisons are employed. Because of the nature of the data from NRTs and CRTs, the way to use the data from each will be a bit different. The approach to NRT data will be presented first. It will then be modified to fit CRT data.

In presenting data to determine weaknesses or strengths, some indication of the error in each test score should be considered along with the absolute test scores. The inclusion of test error is needed to prevent drawing the conclusion that a school is weak in math because its score in that area is two points below its score in reading. A good metric to use here is normal curve equivalent (NCE) scores. Their equal-interval quality is needed to look at score differences. Additionally, they will have the same meaning for all subtests in a battery. Other equal-interval metrics, such as expanded scale scores, are not appropriate as they do not have the same meaning for all subtests. A standard should be set to determine meaningful differences so that schools have some guidelines as to when special action will be needed. The guideline may be determined using traditional tests of statistical significance. However, the problem discussed earlier of small differences being statistically significant in large groups can apply here also. Given this situation, it may make more sense to specify

some amount of difference that appears to make intuitive sense. The standard can be modified if it seems to over- or underidentify problem areas.

Group results for a CRT are generally a report of the percentage of students passing each objective. A comparison of these percentages passing on all objectives can be made just like the comparison of NCE means described above for NRTs. However, to compare the percentages passing each objective assumes that the objectives are both of equal difficulty and are covered equally well by the curriculum. If this is not the case, it will be necessary to determine if the differences in difficulty are caused by an underlying skill hierarchy, by sloppy test construction, or by weaknesses in the instructional program.

If results on different CRTs (e.g., reading and math) are being compared, the caution presented earlier must be dealt with. That is, there may be different standards on the two tests. To determine if this is the case, you might choose a NRT that measures both subjects and see if the results on that test are similar to those on the CRTs. If not, the reason could be different standards. The recommendation that the comparison test be a NRT is made because those results are not dependent on standard setting.

Once the statistical operations described above have identified areas of weakness or strength, the description of test content discussed in the previous chapter can be used to help a school or district take action. A list of the specific objectives included on the subtest can be very helpful in isolating the skills that need to be improved or those that are being taught very well.

DETERMINING IF A SCHOOL DID AS WELL AS IT SHOULD HAVE

Many people are aware of at least some of the reasons students or groups of students do not all perform the same on achievement tests. Thus, when scores for a given school are not at the top of the distribution, the natural question is often, did the school do as well as it should have done? We have no easy, incontrovertible way to respond to that question. Some people argue it is simply a matter of administering an abilities test, an achievement test, and then comparing the results of the two tests to answer the question. However, this premise that there are group-administered tests which measure something called "ability" which can be distinguished from "achievement" has been severely challenged. Standardized, group-administered ability tests usually assess skills in reading, computing, and other areas that are learned and which strongly resemble the skills assessed on achievement tests. Thus, using the performance on one as a standard against which to measure the other is highly questionable.

Given these real limitations in our measuring instruments, this question cannot be answered absolutely. As an alternative, the question which might be asked is whether a school is making appropriate progress. To address this issue, one can use past achievement test scores to predict future ones as described in the previous discussion of longitudinal analysis. For NRTs, percentile ranks (or NCEs which are directly related to percentile ranks) would be a good metric to use for this purpose. Prediction of performance is based on the following assumption: if a group averages at the 85th percentile in Grade 3, it is

expected that the same group would perform at about the same percentile level in succeeding grades, if normal progress were made. Deviations in either direction indicate that something unusual is occurring. Again, establishing a guideline for the point when a deviation becomes large enough to be important must be based on professional judgment and practical experience.

It should be pointed out that for an analysis such as the one described above, choice of a metric is critically important. For example, grade equivalent scores would not be appropriate because it would be extremely difficult to define normal growth from them. Students at the 50th percentile might be expected to improve by 1 year for each year in school. However, a student at the 80th percentile may improve, depending on the grade, anywhere from 1.5 to 3 or more years in a school year. Those at the 20th percentile may be doing well to improve .6 of a year in that time.

On CRTs, percentile rank may be replaced by the number of objectives passed at two points in time. Success at the second data point would be determined by whether the school had achieved more or less objectives than did the typical school which started with the same number achieved. Determining how many more or less objectives passed constitute a warning signal is a decision which must, again, be left to professional judgment.

Chapter 9

SUMMARY

Not too long ago, test results were considered the private domain of teachers, counselors, and other school staff members. In the past ten years, with the strong educational accountability movement, this is no longer the case. The present task for the school district test director is to communicate test results, not to make sure they remain confidential. In this paper, we have tried to provide guidelines for how this might be accomplished. We have discussed the contents of test reports, and how the approach to reporting might be modified to meet the needs of different audiences.

REPORT CONTENTS

We have divided our discussion of report contents into three major areas: Descriptive Information, Test Results, and Interpretive Cautions. The descriptive information includes a description of the test program such as names of test batteries, grades tested, and dates of administration. A discussion of the skills measured by each subtest is also important. Finally, this section should provide an explanation of the types of scores that are used in the report.

We recommend that the reporting of test results include a measure of the average performance of the groups of interest--district, school, special programs, etc. In addition, some indication of the dispersion of scores in the group should be provided. One way of doing this is by

showing the percentage of students that scored in each national quarter of the national norm group. Historical data should also be included to provide a picture of whether the achievement level in a school or district is improving or declining. Additional data that can be helpful to a district in planning instructional programs are results by racial/ethnic group, by sex, or by groups of students with similar socioeconomic status.

We also recommend that test reports clearly explain the limitations of test scores. Without such an explanation (and, unfortunately, sometimes even with it) people will almost assuredly misuse the results. Areas of special concern include the interpretation of grade equivalent scores; the comparison of performance across tests, grades, schools, or groups of students; and the interpretation of small changes in performance.

Many of the elements that we have suggested be included in reports of test data are also mentioned in the Standards for Educational and Psychological Tests, published by the American Psychological Association (APA, 1974). One of the areas mentioned in that document is that the influence of race, sex, and socioeconomic status on test performance should be pointed out. The Standards also call for warning against common misuses of test scores and for providing sufficient information for correct interpretation.

AUDIENCES

Three audiences were discussed here: the board of education (and the general public), parents, and school-based staff. While the information

needs of these audiences were judged to be similar in many aspects, it was recommended that reports be somewhat differentiated in terms of comprehensiveness and format.

Reports to boards of education (and the public) are generally the most formal and complete, including (where poseible) most of the information reviewed above. Reports to parents are generally much briefer and deal with a child's performance, not with group data. These reports might contain a couple of sentences describing the program and a nontechnical explanation of the results or of how to interpret the data that are presented. More critical is a clear discussion of test error, since parents often feel that a 1 or 2 percentile rank change is a meaningful trend. Reporting scores with error bands can help in getting across the idea of test error.

School-based staff members can probably use the most detailed reports on test data for their own school. This report need not, however, be as formal as the annual report presented to the school board. Often these reports are in the form of printouts with little accompanying text. This is because the staffs generally receive the same kind of reports each year and may not require much explanation after an initial workshop. These reports can include detailed frequency distributions; results for students grouped by class, score, or special program; and school historical trends.

A FINAL WORD

Looking over these chapters and the materials received from school districts, we feel compelled to ask, "How did something as simple as

reporting test scores get so complicated?"

We use printouts, bar graphs, pie graphs, tables, exhibits, brochures, overlays, and slide/tapes. We have formal reports, summary reports, conferences, and workshops.

We could probably have doubled the length of this discussion had we singled out the "press conference" for additional attention. Undoubtedly, a summary of the approaches used in this area would comprise a valuable, and amusing, volume of its own.

While the "art" of reporting test scores certainly can be improved, we know of no way to drastically streamline the task of reporting. There currently exists no all-purpose approach which can be adopted for all audiences and all districts, nor do we feel that one is likely to emerge in the near future. The needs of each group must be kept in mind and the format and content of each report modified accordingly. The critical thing is to keep in mind the question(s) that each audience needs to have answered and to provide the information which will allow accurate interpretation of the answers provided.

If there is one point that we cannot make strongly enough, it is that reports to each and every audience must be structured to answer questions, not just provide numbers. This means that the knowledge base, concerns, and experience of an audience need to be considered very carefully. For some, this means printouts; for others, a simple letter. There is no longer any question about whether test scores should be released. The "how to do so" remains that which each of us must solve.

REFERENCES

American Psychological Association. Standards for Educational and Psychological Tests. Washington, DC: American Psychological Association, 1974.

Educational Research Service. Releasing Standardized Achievement Test Scores to the Public. Arlington, VA: Educational Research Service, 1974.

National School Public Relations Association. Releasing Test Scores: Educational Assessment Programs, How to Tell the Public. Arlington, VA: National School Public Relations Association, 1974. (ERIC Document Reproduction Service No. ED 119 322).

APPENDIX A

TEST RESULT REPORTING MATERIALS REVIEWED

District	Report
Albuquerque	Comprehensive Tests of Basic Skills (CTBS) Testing, Spring 1982, District Report
Austin	Individual Student Reports Student Achievement, 1982-83
Charleston County	Report to Parents Results from the Spring, 1982, Norm-Referenced Testing Program
Dade County	District and School Profiles, 1982-83 Parent Report
Dallas	An Interpretive Analysis of System-Wide Achievement Data, 1981-82 Do You Know About Testing? - Topics for Parents School Achievement Indices
Detroit	Student Test Report Summary of Achievement Test Scores, 1982
District of Columbia	A Summary of Student Achievement on the Comprehensive Tests of Basic Skills
Fort Worth	Press release
Houston	Elementary School Profiles, 1981-82 Secondary School Profiles, 1981-82
Los Angeles	Report on the District Testing Programs, 1981-82 Norm-referenced Test Results, 1981-82
Memphis	California Achievement Tests, A Practical Guide for Using and Interpreting the Results
Montgomery County	Annual Test Report, 1979-80 Annual Test Report, 1981-82
New Orleans	Testing Programs 1981-82, Summary of Results and Interpretive Guide
Palm Beach County	Sample School Report

Pittsburgh	Preliminary Report on Student Achievement in the Pittsburgh Public Schools, School Year 1981-82 Report to Parents
Portland, OR	General Orientation Manual for the Portland Public School Achievement Testing Program Portland Public Schools Achievement Levels Tests, Sample Reports
Rochester	Elementary School Profiles for Academic Year 1982-83
San Diego	California Assessment Program Statewide Testing Results by District and by School, 1981-82 School Year Districtwide Testing Results by District and by School, Grade 11, Fall 1982

APPENDIX B

COMMONLY USED TEST TERMS

This appendix provides information about commonly used test terms. Each term is defined. The definition is followed by a statement on its uses and a list of precautions to be observed when using the type of test or score being discussed. The terms are listed in alphabetical order.

CRITERION-REFERENCED TEST (CRT)

Definition

A test based on specific learning objectives (or teaching objectives), usually within a narrow range of subject matter or skill. The tests are designed to measure the knowledge or skills the student has attained. The Maryland Functional Reading Test (MFRT) is an example of a CRT.

Use

CRTs provide information about the extent to which the student has attained the learning objective(s).

Precautions(s)

1. CRTs are often designed so a student can answer all or almost all of the questions correctly or incorrectly depending on the extent to which the student has attained the skills being measured. They are not designed to yield information about different levels of achievement and, therefore, cannot usually be used to rank students on specific skills.

2. To be useful measures of specific skills, CRTs must have a sufficient number of questions measuring each particular skill included on the test. Although what is "sufficient" is not a fixed number, there should, in most cases, be at least five questions which measure a skill. A test purporting to be a CRT which has fewer than five questions per skill should be viewed with skepticism.

GRADE EQUIVALENT SCORES (GE)

Definition

The grade equivalent of a given raw score on any test estimates the grade level at which the typical pupil achieves this raw score. The digit(s) to the left of the decimal point represent the grade; the digit to the right of the decimal point represents the month within the grade according to the following table:

Number	Month
0	September
1	October
2	November
3	December
4	January
5	February
6	March
7	April
8	May
9	June-August

An example of how a test publisher might derive grade equivalents can be useful in understanding GE. The example presented below represents the best methodology currently in use. Many tests are normed with fewer samples.

If the publisher is norming a fourth grade test, he will test a representative sample in Grades 3, 4, and 5. In each grade, the sample, or two comparable samples, will be tested in the fall (November) and the spring (April). Thus, the grade levels being tested as 3.2, 3.7, 4.2, 4.7, 5.2, and 5.7. (Often publishers test only once a year.)

The average raw test score for the students in each group is computed and plotted on a graph similar to the one below. The mean scores are indicated by points on the graph. All other grade-and-month values are estimated by interpolation between the means and extrapolation beyond the means. The GEs beyond the grade range of students in the norming sample should be regarded as no better than rough estimates.

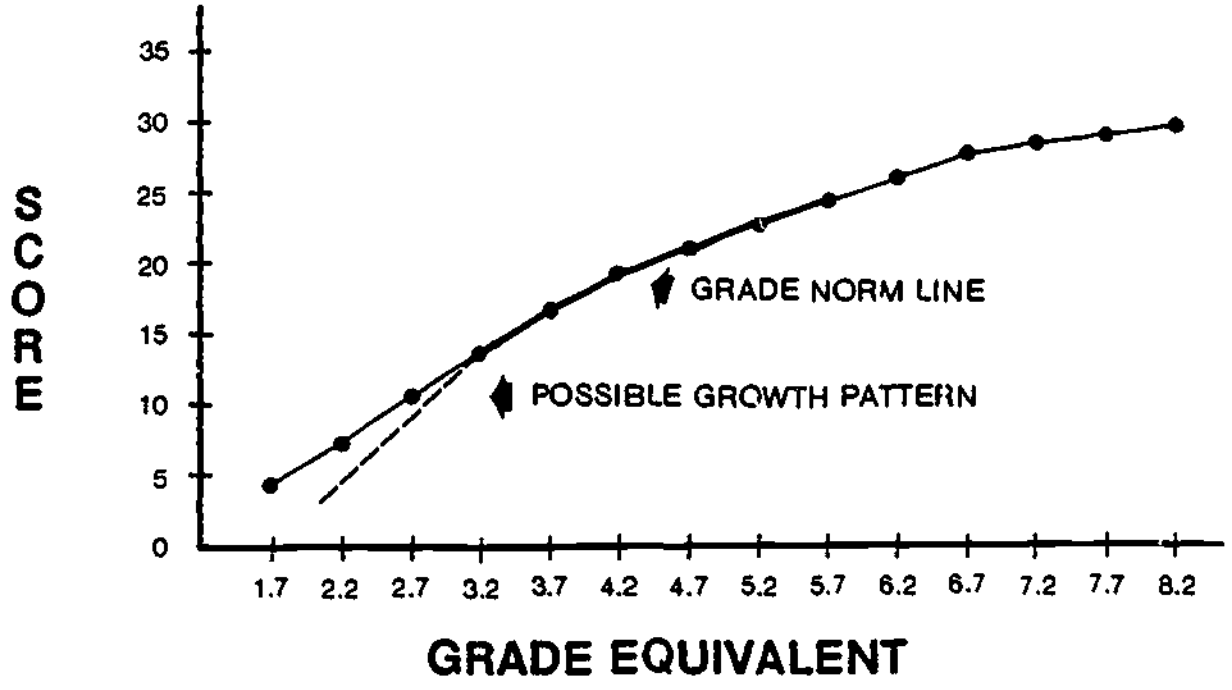


Figure B1

Use

GEs provide a familiar referent for test scores.

Precautions

1. The grade equivalent score does not indicate the grade level of work that a student can perform. It simply estimates the grade level of the typical student in the norming sample achieving a given raw score. For example, suppose a fourth grade student has a score with a grade equivalent of 5.4 on a fourth grade test. This does not mean that a fourth grade student can do work which is done in January in the fifth grade. It simply estimates that this student did as well on a fourth grade test as the typical student in January of the fifth grade. However, remember that if the norming sample for the fourth grade test did not include any fifth grade students, this estimate is very tentative.

2. Grade equivalent scores should not be added and subtracted, because they are not an equal distance apart at all points. They are developed under an assumption that learning occurs equally during the school year. In fact, students tend to learn more at different times in the year. From a strict statistical point of view, this lack of equal score intervals means that mean GE scores should not be computed. However, if the GE scores are converted to Normal Curve Equivalent scores which do have this equal interval quality, the mean score computed from the converted scores is generally very close to that computed from the GEs, especially if the grade equivalents represent a wide range of possible scores.

3. The attempt to build a scale based on the assumption of equal learning cited in Number 2 above results in differential GE gains for raw score changes. What occurs is that a one raw score point change may cause a one-month change in GE at one place in the norm table and a five-month gain elsewhere. The largest changes in GE generally happen in the extremes of score distribution.

An example of the unequal GE differences between raw scores is shown below. These scores are taken from the Iowa Tests of Basic Skills (ITBS) seventh grade spelling test.

Grade	Test	Raw Score	Grade Equivalent	Difference in Grade Equivalent
7	Spelling	7	3.5	
7		8	4.0	.5
7		9	4.4	.4
7	Spelling	25	8.4	
7		26	8.5	.1
7		27	8.7	.2

4. Grade equivalents generally have a wider range at higher grade levels. This leads to the situation in which a student who has the same PR in Grades 3 and 5 will probably be farther above (or below) the median in GE terms in Grade 5. This means that if he/she has a high PR in both grades, the gain in GE terms will be more than two years. If he/she has a low PR, the gain will be less than two GEs. Therefore, if a constant expected GE gain were established for all students, it would be too high for some and too low for others. The example below from ITBS norms demonstrates this problem.

PR	Grade 3	Grade 5	Grade Equivalent Change
90	5.1	7.5	2.4
50	3.6	5.6	2.0
10	2.6	4.1	1.5

5. Because a grade equivalent score represents the performance of a typical student at a given grade level, approximately half of the students in a nationwide sample would be expected to score below grade level.

6. Grade equivalents should not be compared across subject areas, because they have different meaning. For example, mathematics is more grade-related than reading; therefore, the GEs are generally less spread out for math than for reading.

7. Grade equivalents should not be compared across different tests because they may have different means because of different norming samples.

INTERQUARTILE RANGE

Definition

Quartiles are scores (points in a distribution) that divide a score distribution into quarters. Twenty-five percent of the scores are at

or below the first quartile (Q1), 50 percent are at or below the second quartile (Q2, which is also the median), and 75 percent are at or below the third quartile (Q3). The interquartile range includes the band of scores that lies between Q1 and Q3, or the middle 50 percent of the scores.

Use

By eliminating the effect of the lowest and highest quarters of the distribution, the interquartile range provides a measure of how the typical students in a group performed.

Precaution(s)

Eliminating the extreme scores may be removing important information such as the location of pockets of students needing compensatory or gifted programs. If the median is close to either quartile, it could indicate a large number of students at that end of the distribution who might require such services.

MEAN

Definition

The sum of the scores divided by the number of scores.

Use

The mean is used as measure of the performance of the "typical" student in a group.

Precautions

1. In a small group, the mean can be overly influenced by a few extreme scores. Thus, if a few scores in a distribution are very low but most are quite high, the mean will be depressed by the low scores more than the median. In groups where there are a few extremely low scores, the mean will, therefore, be lower than the median. Therefore, it is often useful to compare the mean with the median.

2. Use of the mean provides no information about the spread of scores.

MEDIAN

Definition

The score that divides a test score distribution in half is known as the median. Half of the scores are above the median, half are below. The median is the score that has a percentile rank of 50.

Use

The median is used as a measure of the performance of the "typical" student in a group.

Precaution(s)

1. See Precaution 1 for "mean."

2. Use of the median provides no information about the spread of scores.

NORMAL CURVE EQUIVALENT SCORES (NCE)

Definition

NCEs divide the normal distribution into 99 segments, units, or scores (Figure B2). Scores range from 1-99, with a mean/median of 50. NCEs can be related to percentile ranks as shown in the comparative scales in Figure B2.

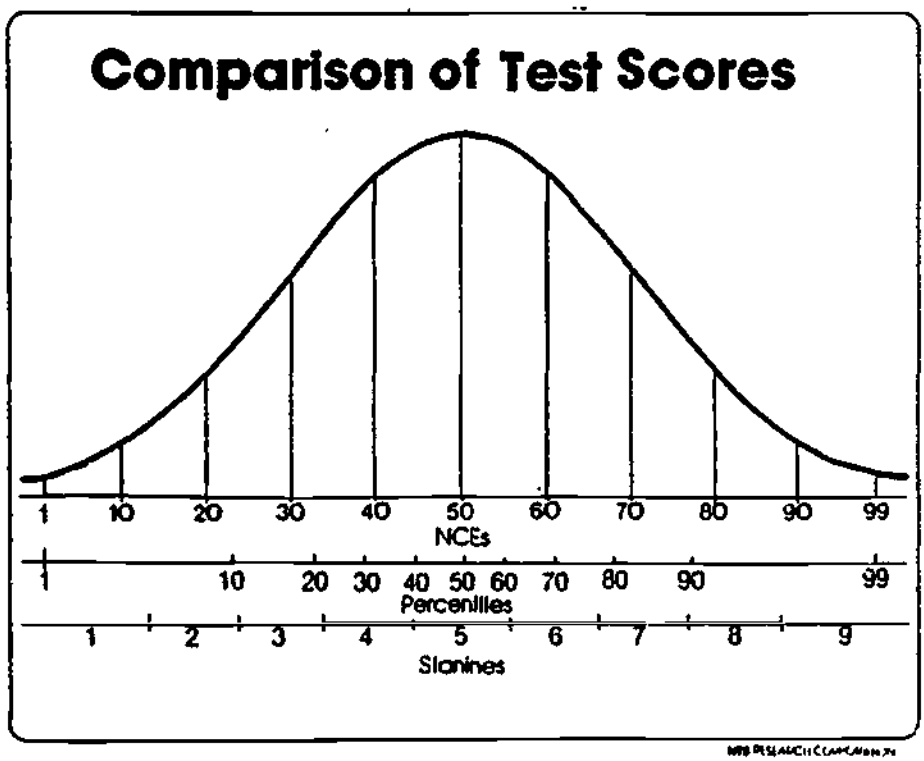


Figure B2

Use

1. NCEs can be subjected to arithmetic operations. Therefore, mean NCEs can be computed, and differences in NCEs can be compared at all points in the score distribution.[1]

2. NCEs can be used in analyses of group data (for reasons above). In addition, NCEs are scaled to reveal small changes, something which stanine scores will not do consistently because of the large score range at each stanine point.

Precaution(s)

1. Use of NCEs for evaluating individualized performance is to be done with caution. A change of five NCE units on a test score is within the error range for individuals on most standardized tests. However, since NCEs give a false sense of precision--and hence of security--the careless test user could consider such a change meaningful.

2. NCEs are difficult to interpret when presented alone. After an analysis has been performed on the basis of NCEs, results are often converted to some more readily understandable scale like percentile ranks.

NORM-REFERENCED TEST (NRT)

Definition

The NRT is designed to rank students according to the number of test items answered correctly (i.e., according to raw score). Ranking is usually also done in relation to the performance of a norming sample. The California Achievement Tests is an example of an NRT.

Use

Norm-referenced tests identify those students who know the most about the content included on the test.

Precaution(s)

1. A good NRT is designed to enable between 40 and 70 percent of the examinees to answer any given item correctly. Many items are therefore too difficult for a majority of examinees to get right. This means that most NRTs are not very good tests of what an individual student knows (as opposed to

---

[1] In a strict statistical sense, it is probably incorrect to subject any test scores to arithmetic operations. However, NCEs, standard scores with an underlying normal distribution, raw scores, and stanines come closer than any other score scales to having equal-interval properties which permit arithmetic operations.

criterion-referenced tests). Rather, they are measures of who knows the most about the test content.

2. NRTs often include only one or two questions which measure achievement of a given skill or objective. Information about student performance on a particular objective is, therefore, usually not very reliable.

PERCENTILE RANK (PR)

Definition

The percentile rank (PR) expresses the percentage of students in the norming sample who scored at or below a given score. For example, if a raw score of 30 has a percentile rank of 78, then 78 percent of the students in the norming sample scored at or below 30 items correct.

Use

PRs provide easily interpretable information about how a given student's performance on a test compares with the performance of students in the norming sample.

Precaution(s)

1. PRs should not be added nor subtracted because they are not an equal distance apart at all points. For example, Figure 3.2 clearly shows that an increase of 10 points between percentile ranks 45 and 55 in not the same distance as an increase of 10 points between percentile ranks 85 and 95. A person would have to show a larger amount of improvement to achieve the second increase.

2. On a test of fewer than 100 questions, it is not possible for every whole number of the percentile rank scale to have an associated raw score. Therefore, in such circumstances, a one-point increase in raw score can cause an increase of several percentile rank units. What might appear to be substantial increase on the percentile rank scale is really only an increase of one additional question correct. This caveat applies to virtually all tests in standardized batteries.

3. Percentile ranks should not be confused with percent of correct answers (raw scores). They have completely different meanings.

RAW SCORE

Definition

Raw score represents the number of questions or test items answered correctly.

Use

Raw scores can be used to report the number of questions answered correctly.

Precaution(s)

1. A raw score has no meaning other than the number of items answered correctly. It provides no interpretative information.

2. Raw scores can be quite misleading when reported by themselves because the meaning of raw scores differs from test to test. For example, if one 50-item test is easy and one 50-item test is difficult, a raw score of 30 on the difficult test might represent better performance than a raw score of 45 on the easier test.

3. Subjecting raw scores to arithmetic operations (e.g. addition, etc.) is a questionable procedure. Generally, raw scores do not have the equal inteval property required for these operations. This is because the same raw score can be obtained by different students who get different combinations of items correct. These items will most likely vary in their level of difficulty. Thus, identical raw scores will possibly represent differential levels of achievement.

STANDARD DEVIATION (SD)

Definition

Standard Deviation (SD) is a measure of the dispersion in a set of scores. The closer the scores cluster around the mean, the smaller the SD will be.

Use

As a measure of the spread in a set of scores, the SD can be used to assist in determining the degree of importance of score differences. For example, a difference of 2 points would probably not have much meaning if the SD were 20 but could be quite important if the SD were 0.5.

STANINE

Definition

A stanine is one of the scores of nine-point division of the normal distribution. Stanine scores range from 1 to 9 with a mean and median of 5. As shown in Figure B2, each stanine has a range of corresponding percentile ranks or raw scores.

Uses

1. Stanines can be subjected to arithmetic operations (addition, etc.). Therefore, the mean of distributions can be computed, and differences in stanine scores can be compared at all points

in the distribution except, in some cases, at the extreme stanine scores of 1 and 9.

2. Stanines do not give a false sense of accuracy of a given score because each stanine covers a range of raw scores. The stanine scale is therefore useful for reporting individuals' scores. Differences in stanines are more likely to represent change beyond that which can be attributed to error than are other kinds of scores.

Precaution(s)

As can be seen in Figure B2, interpretation of differences in stanine scores is clouded by the range within a given stanine. For example, if an individual's score increases from the top of the Stanine-3 range to the bottom of the Stanine-5 range, it represents less improvement than an increase from the bottom of the Stanine-3 range to the top of the Stanine-4 range. However, on cursory examination, it would seem as if the first increase were the greater.

APPENDIX C

TESTING TEXTBOOKS THAT INCLUDE DISCUSSIONS OF TESTING TERMS

Anastasi, Anne. Psychological Testing. Macmillan Publishing Co., New York, N.Y., 1982

Cronbach, Lee J. Essentials of Psychological Testing. Harper & Row, New York, N.Y., 1970

Ebel, Robert L. Essentials of Educational Measurement. Prentice-Hall, Englewood Cliffs, N.J., 1979

Hopkins, Kenneth D., and Stanley, Julian C. Educational and Psychological Measurement and Evaluation. Prentice-Hall, Englewood Cliffs, N.J., 1981

Mehrens, William A., and Lehmann, Irvin J. Measurement and Evaluation in Education and Psychology. Holt, Rinehart and Winston, Inc., New York, N.Y., 1973

Thorndike, Robert L., and Hagen, Elizabeth P. Measurement and Evaluation in Psychology and Education. John Wiley & Sons, New York, N.Y., 1977

APPENDIX D
REPORTS OF TEST RESULTS
cited in
"RESEARCH AND EVALUATION STUDIES
FROM LARGE SCHOOL DISTRICTS 1982"*

ALBUQUERQUE PUBLIC SCHOOLS, NEW MEXICO

New Mexico High School Proficiency Examination. Spring, 1980 Test Results. Albuquerque Public Schools, 1980. (ERIC Document Reproduction Service No. ED 211 563).

ATLANTA INDEPENDENT SCHOOL DISTRICT, GEORGIA

McCarson, Carole. Reading Achievement. Report No. 14-8. Atlanta Public Schools, June 1980. (ERIC Document Reproduction Service No. ED 210 665).

McCarson, Carole. Results of the Admissions Testing Program for the Atlanta Public Schools' Seniors from 1975 to 1981. Atlanta Public Schools, February 1982. (ERIC Document Reproduction Service No. ED 217 068).

McCarson, Carole. Results of the Georgia Statewide Testing Program for the Atlanta Public Schools, 1981. Atlanta Public Schools, Division of Research, Evaluation, and Data Processing, 1981. (ERIC Document Reproduction Service No. ED 217 067).

AUSTIN INDEPENDENT SCHOOL DISTRICT, TEXAS

Austin Independent School District Achievement Profiles, 1980-81. Volume I: Elementary Schools (Iowa Test of Basic Skills), Allan-Linder and District Publication No. 80.83. Austin Independent School District, Office of Research and Evaluation, June 30, 1981. (ERIC Document Reproduction Service No. ED 209 290).

Austin Independent School District Achievement Profiles, 1980-81. Volume II: Elementary Schools (Iowa Tests of Basic Skills), Maplewood-Zilker. Austin Independent School District, Office of Research and Evaluation, 1981. (ERIC Document Reproduction Service No. ED 209 291).

* This bibliography, and earlier annual editions (1980,1981), are available from the ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, NJ 08541-0001, for $6.00 each.

Austin Independent School District Achievement Profiles, 1980-81. Volume III: Junior High Schools (Iowa Tests of Basic Skills) and Senior High Schools (Sequential Tests of Educational Progress). Austin Independent School District, Office of Research and Evaluation, June 30, 1981. (ERIC Document Reproduction Service No. ED 209 292).

DETROIT PUBLIC SCHOOLS, MICHIGAN

Summary of Achievement Test Scores--1980. School-by-School Test Results. Detroit Public Schools, Department of Research and Evaluation, 1980. (ERIC Document Reproduction Service No. ED 208 051).

PHILADELPHIA CITY SCHOOL DISTRICT, PENNSYLVANIA

Grosswald, Jules. City-Wide Summaries, City-Wide and District Performance Distributions, Kindergarten through Grade Twelve. 1978-79 Philadelphia City-Wide Testing Program, February 1979 Achievement Testing Program. Report No. 8004. Philadelphia School District, Office of Research and Evaluation, September 1979. (ERIC Document Reproduction Service No. ED 208 052).

SAN DIEGO UNIFIED SCHOOL DISTRICT, CALIFORNIA

Statewide and Districtwide Testing Results by District and by School, San Diego City Schools. December 1979 to October 1980. San Diego City Schools, November 1980. (ERIC Document Reproduction Service No. ED 212 641).

Testing Results for Minority Isolated Schools. San Diego City Schools. Spring 1981. Report No. 295. San Diego City Schools, Evaluation Services Department, July 7, 1981. (ERIC Document Reproduction Service No. ED 210 335).

ERIC/TM Report 85

REPORTING TEST SCORES TO DIFFERENT AUDIENCES

by

Joy A. Frechtling

and

N. James Myerberg

December 1983

Ten years ago, the practice of releasing test scores to the public was not generally accepted. The issue today is not whether or not to release test scores, but rather what to release and how to release it. Further, it has been increasingly acknowledged that since the audience for test scores has different faces with different backgrounds or interests, the content and format of reporting may also need to be varied.

The purpose of this report is to address issues in the release of test scores to a variety of audiences: parents, school board members, school staff, the news media, and the general public. It discusses the kinds of information that such reports might include and suggests some strategies for presenting them.

****

ORDER FORM

Please send_____copies of ERIC/TM Report 85, "Reporting Test Scores to Different Audiences," at $7.00 per copy.

Total Encloaed $_____

Name_____

Address_____

_____Zip_____

Return this form to:
ERIC/TM
Educational Testing Service
Princeton, NJ 08541-0001

RECENT TITLES
IN THE ERIC/TM REPORT SERIES

#84 - Assessment of Learning Disabilities, by Lorrie A. Shepard. 12/82 $6.50.

#83 - Statistical Methodology in Meta-Analysis, by Larry V. Hedges. 12/82, $7.00

#82 - Microcomputers in Educational Research, by Craig W. Johnson. 12/82, $8.50.

#81 - A Bibliography to Accompany the Joint Committee's Standards on Educational Evaluation, compiled by Barbara M. Wildemuth. 12/81, $8.50.

#80 - The Evaluation of College Remedial Programs, by Jeffrey K. Smith and others. 12/81, $8.50.

#79 - An Introduction to Rasch's Measurement Model, by Jan-Eric Gustafsson. 12/81, $5.50.

#78 - How Attitudes Are Measured: A Review of Investigations of Professional, Peer, and Parent Attitudes toward the Handicapped, by Marcia D. Horne. 12/80, $5.50.

#77 - The Reviewing Processes in Social Science Publications: A Review of Research, by Susan E. Hensley and Carnot E. Nelson. 12/80, $4.00.

#76 - Intelligence Testing, Education, and Chicanos: An Essay in Social Inequality, by Adalberto Aguirre Jr. 12/80, $5.50.

#75 - Contract Grading, by Hugh Taylor. 12/80, $7.50.

#74 - Intelligence, Intelligence Testing and School Practices, by Richard DeLisi. 12/80, $4.50.

#73 - Measuring Attitudes Toward Reading, by Ira Epstein. 12/80, $9.50.

#72 - Methods of Identifying Gifted Minority Students, by Ernest M. Bernal. 12/80, $4.50.

#71 - Sex Bias in Testing: An Annotated Bibliography, by Barbara Hunt. 12/79, $5.00.

#70 - The Role of Measurement in the Process of Instruction, by Jeffrey K. Smith. 12/79, $3.50.

#68 - The Educational Implications of Piaget's Theory and Assessment Techniques, by Richard DeLisi. 12/79, $5.00.

#65 - The Practice of Evaluation, by Clare Rose and Glenn F. Nyre. 12/77, $5.00.

#63 - Perspectives on Mastery Learning & Mastery Testing, by Jeffrey K. Smith. 1977, $300.