DOCUMENT RESUME

ED 208 061 TM 810 832

TITLE Accountability Testing Handbook.
INSTITUTION Montgomery County Public Schools, Rockville, Md.
PUB DATE Aug 80
NOTE 77p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Definitions; Elementary Secondary Education; Objectives; Scores; *Standardized Tests; *Test Format; *Testing; *Test Interpretation
IDENTIFIERS *California Achievement Tests; Montgomery County Public Schools MD; Test Reporting

ABSTRACT

The purpose of this handbook is to acquaint principals and teachers with the California Achievement Tests, mandated by the Maryland State Department of Education. Reports of the test results are also discussed. The first chapter describes the test and provides examples of question formats. A table of the objectives measured is also included. The second chapter presents reports that are distributed to the schools, and an explanation of the data on the reports with suggestions for their use. Also included are the School Frequency Distributions, Mean Score Report, Percent Correct by Objective, and the Individual Test Report. Commonly used technical testing terms are defined in the final chapter. (Author/GK)

MONTGOMERY COUNTY PUBLIC SCHOOLS

ACCOUNTABILITY TESTING HANDBOOK

August 1980

ACCOUNTABILITY

TESTING

HANDBOOK

Department of Educational Accountability
MONTGOMERY COUNTY PUBLIC SCHOOLS
Rockville, Maryland

TABLE OF CONTENTS

INTRODUCTION

The purpose of this handbook is to acquaint principals and teachers with the California Achievement Tests, the standardized test mandated by the Maryland State Department of Education. Reports of the results from the California Tests are also discussed in order to assist staff in interpreting them accurately and completely.

Chapter 1 describes the test and provides examples of question formats for the purpose of familiarizing teachers with the types of questions and directions used throughout the subtests. Also included in this chapter is a table of the objectives measured by each subtest at levels 13-19.

Chapter 2 presents reports that are distributed to the schools, an explanation of the data on the reports, and suggestions for accurate use of these data. The School Frequency Distributions give schools a picture of both their typical achievement and the variation of achievement in the school. The Mean Score Report summarizes the typical achievement in each school. A discussion of longitudinal and nonlongitudinal analyses describes how to overcome some weaknesses of the Mean Score Report. The "Percent Correct by Objective" report provides school, area, and county summary information on the objectives measured by each subtest. Finally, the "Individual Test Report" provides data on individual students.

The final chapter presents the more commonly used technical testing terms with definitions, as well as some important precautions about their use.

CHAPTER 1

CALIFORNIA ACHIEVEMENT TESTS

The major purpose of this chapter is to describe the characteristics of the California Achievement Tests, to provide a list of the skills measured by each subtest and to provide examples of question formats used in the California subtests, Levels 13 to 19. The purpose is not to inform teachers of what the test items are. The examples used are meant to acquaint teachers with the types of questions used and alert them to the need for careful attention to directions. There is no intent to "teach to the test" nor to give last minute training to students. Finally, the examples are not meant as indications of what has to be taught.

1A. DESCRIPTION OF CALIFORNIA ACHIEVEMENT TESTS

The California Achievement Tests replace the Iowa Tests of Basic Skills (ITBS) and the Tests of Academic Progress (TAP) as the standardized norm-referenced[1] test used for systemwide testing. There are a few basic changes, as well as many similarities.

There are five major content areas measured on the California Achievement Tests, shown in Table 1.1. They are Reading, Spelling, Language, Math and Reference Skills. Most levels of the California Tests measure the same content areas as the ITBS. However, some areas measured on separate subtests by the ITBS have been combined into one subtest on the California Tests. This occurs with Punctuation and Capitalization, measured separately on the ITBS, but included in the same subtest on the California. In addition, the three ITBS subtests dealing with Reference Skills have been combined into one subtest on the California. On the other hand, Mathematics Computation, a subtest on the California, is not directly measured on the ITBS.

---

[1] Norm-referenced is explained in Chapter 3.

TABLE 1.1

COMPARISON OF SUBTESTS ON THE CALIFORNIA ACHIEVEMENT TESTS (Levels 13-19) AND IOWA TESTS OF BASIC SKILLS (ITBS)

CALIFORNIA	ITBS
LEVEL 13 ONLY: PHONIC ANALYSIS	
LEVEL 13 ONLY: STRUCTURAL ANALYSIS	
READING VOCABULARY	VOCABULARY
READING COMPREHENSION	READING COMPREHENSION
SPELLING	SPELLING
LANGUAGE MECHANICS	PUNCTUATION; CAPITALIZATION
LANGUAGE EXPRESSION	LANGUAGE USAGE
MATHEMATICS COMPUTATION	
MATHEMATICS CONCEPTS AND APPLICATIONS	MATHEMATICS CONCEPTS; MATHEMATICS PROBLEM SOLVING; GRAPHS AND TABLES
REFERENCE SKILLS	REFERENCE SKILLS; MAP READING; GRAPHS AND TABLES

Level 13 of the California, which can be used in Grade 3, has two additional reading sections, Phonic Analysis and Structural Analysis. It does not have a Reference Skills section.

The content covered in the California Tests does not match as closely with the TAP (given in Grade 11) as it does with the ITBS, shown in Table 1.2. The California Tests include three content areas not measured by the TAP. These are Reading Vocabulary, Mathematics Computation, and Reference Skills. However, Social Studies, Science, and Literature, all measured by the TAP, are not covered in the California Test Battery.

Like the ITBS, the California Achievement Tests include several total scores which are combinations of subtest scores. Total Reading is a combination of the following subtests: Phonic Analysis (Level 13 only), Structural Analysis, (Level 13 only), Reading Vocabulary and Reading Comprehension. Total Language is made up of the Language Mechanics and Language Expression subtests. Spelling is not included in the Language Total. There is also a Total Mathematics score that is composed of the Mathematics Computation and the Mathematics Concepts and Applications subtests. Finally, the Total Battery score is a combination of all of the above subtests. The Reference Skills subtest (Levels 14 to 19) is not included in the Total Battery score.

TABLE 1.2

COMPARISON OF SUBTESTS ON THE CALIFORNIA ACHIEVEMENT TESTS (Levels 13-19) AND TESTS OF ACADEMIC PROGRESS (TAP)

CALIFORNIA	TAP
READING VOCABULARY	READING
READING COMPREHENSION	
SPELLING	ENGLISH
LANGUAGE MECHANICS	
LANGUAGE EXPRESSION	
MATHEMATICS COMPUTATION	MATHEMATICS
MATHEMATICS CONCEPTS AND APPLICATIONS	
REFERENCE SKILLS	SOCIAL STUDIES
	SCIENCE
	LITERATURE

1B. SKILLS MEASURED

Table 1.4 shows the skills measured by each subtest of the California Achievement Tests, Levels 13 to 19. More detailed descriptions of the objectives on each level can be found in the Class Management Guide for the California published by CTB/McGraw-Hill. A copy will be available in each school.

The test levels are recommended for administration by the publisher at grade ranges that overlap. The levels and recommended grade ranges are shown in Table 1.3.

TABLE 1.3

TEST LEVELS AND GRADE RANGES

Level	Range
13	2.6 - 3.9
14	3.6 - 4.9
15	4.6 - 5.9
16	5.6 - 6.9
17	6.6 - 7.9
18	7.6 - 9.9
19	9.6 - 12.9

The Montgomery County Public Schools will give Level 13 in Grade 3, Level 15 in Grade 5, Level 18 in Grade 8 and Level 19 in Grade 11.

Table 1.4

CATEGORY OBJECTIVES BY LEVEL FOR
THE CALIFORNIA ACHIEVEMENT TESTS

Test/Category Objectives	Level						
	13	14	15	16	17	18	19
Phonics Analysis							
Consonant Clusters/Digraphs	x						
Short, Long Vowels/ Vowel Combinations	x						
Diphthongs	x						
Variant Vowels/Vowel Combinations	x						
Structural Analyses							
Compound Words/Syllables/ Contractions	x						
Base Words/Affixes	x						
Reading Vocabulary							
Same Meaning	x	x	x	x	x	x	x
Opposite Meaning	x	x	x	x	x	x	x
Multimeaning	x	x	x	x	x	x	x
Reading Comprehension							
Recall of Facts	x	x	x	x	x	x	x
Inferred Meaning	x	x	x	x	x	x	x
Character Analysis	x	x	x	x	x	x	x
Figurative Language	x	x	x	x	x	x	x
Author Attitude/Position		x	x	x	x	x	x
Techniques of Persuasion				x	x	x	x
Real/Unreal Elements	x						
Spelling							
Consonant Phonemes/Graphemes	x	x	x	x	x	x	x
Vowel Phonemes/Graphemes	x	x	x	x	x	x	x
Morphemic Units	x	x	x	x	x	x	x

Table I.4 (Continued)

Test/Category Objectives	Level						
	13	14	15	16	17	18	19
Language Mechanics							
Capitalization of I/Proper Nouns	x						
Capitalization of I/ Proper Nouns/Adjectives		x	x	x	x	x	x
Capitalization of Beginning Words/Titles	x	x	x	x	x	x	x
Punctuation of End Marks	x	x	x	x			
Punctuation of End Marks/Colon Semicolon					x	x	x
Punctuation of Comma	x	x	x	x	x	x	x
Punctuation of Quotation Marks		x	x	x	x	x	x
Language Expression							
Pronouns	x	x	x	x	x	x	x
Verbs	x	x	x	x	x	x	x
Adjectives	x	x	x	x			
Subjects/Verbs	x	x	x	x	x	x	x
Modifying Words	x						
Modifying/Trans. Words		x	x	x	x	x	x
Complete/Incomplete/Run-on		x	x	x			
Verbosity/Repitition					x	x	x
Misplaced Modifiers/Nonparallel					x	x	x
Paragraph Sequence		x	x				
Paragraph Sequence/Topic Sentence				x			
Paragraph Sequence/Topic, Concluding Sentence					x	x	x
Mathematics Computation							
[illegible]ition	x	x	x	x	x	x	x
[illegible]raction	x	x	x	x	x	x	x
Multiplication	x	x	x	x	x	x	x
Division	x	x	x	x	x	x	x

Table 1.4 (Continued)

Test/Category Objectives	Level						
	13	14	15	16	17	18	19
Math Concepts and Application							
Numeration	x	x	x	x	x	x	x
Number Theory	x	x	x	x	x		x
Number Theory/Sentences						x	
Number Sentences		x	x	x			
Number Sentences/Properties	x				x		x
Number Properties		x	x	x			
Common Scales	x	x	x				
Geometry	x	x			x	x	x
Measurement	x	x			x	x	x
Geometry/Measurement				x			
Functions and Graphs					x	x	x
Graphs	x	x		x			
Geometry/Measurements/Graphs			x				
Story Problems	x	x	x	x	x	x	x
Reference Skills							
Title Page/Copyright Page		x					
Table of Contents		x					
Index		x					
Dictionary Page		x	x	x	x	x	x
Map		x	x	x	x	x	x
Table			x	x	x	x	
Library Catalog Cards			x	x	x		
Diagram				x	x		
Form						x	x
Readers' Guide						x	x

1C. QUESTION FORMATS

This section provides examples of test questions in each subtest, organized by test levels.

The question formats change among subtests and sometimes within a subtest. The latter case is especially important to note because, in many cases, students cannot be given new verbal instructions when a format changes within a subtest. Another factor to note is that the questions for some subtests of Level 13 must be read aloud to the students. Test administrators must read these questions very carefully and say no more than the Examiner's Manual requires. In reading this chapter, one should be aware of the level to be used with a specific class because of the changes noted above.

The format examples presented here generally use easy questions and do not reflect the level of the questions on the test. The correct answer for each example is indicated by an asterisk (*).

Phonic Analysis

Level 13

The following two formats are used for the subtest, Phonic Analysis, given only on this level.

1. The student is to find the word that has the same beginning (or ending) as the word given by the teacher.

   Example: The teacher says "shy.. . . shy"

   ○ steeple

   ⊛ ship

   ○ scrap

2. The student should read the word with the underlined part and then choose the word with the same vowel sound.

   Example:

   tree ○ trip ⊛ deep ○ tray ○ die

Structural Analysis

Level 13

The following six formats are used on the subtest, Structural Analysis, given only on this level.

1. The student is given a word and asked to find the word from the list that could be combined with the first word to make another word.

Example:

up — road ○ city ○ school ○ stairs (*)

2. The student is asked to count the number of syllables in the word given on the left.

Example:

doctor — 1 ○ 2 (*) 3 ○ 4 ○

3. The student is given an underlined word and asked to choose the word pair that means the same thing.

Example: we're

- ○ we will
- ○ we were
- ○ we did
- (*) we are

4. The student is asked to choose the word whose underlined part is the base word.

Example:

(*) darker
( ) smallest
( ) undone
( ) calling

5. The student is asked to choose the word that has the prefix underlined.

Example:

( ) continue
( ) office
(*) redo
( ) apart

6. The student is asked to select the word that has the suffix underlined.

Example:

( ) careful
( ) readable
(*) friendly
( ) brighten

Levels 13-19.

The following three formats are included on these levels.

1. The student is to choose the word with the same meaning as the underlined word in the phrase.

   Example: choose the answer

   a. take
   *b. select
   c. read
   d. write

2. The student must choose the word with the opposite meaning of the underlined word in the phrase.

   Example: an interesting book

   a. long
   *b. boring
   c. good
   d. green

3. The student is asked to choose one of three sentences which uses the underlined word the same way as in the given definition.

   Example: suddenly, or briskly

   a. She closed the snap on her jacket.
   b. He began to snap at his friend.
   *c. The cord broke with a snap.

Reading Comprehension

Level 13

The following formats are used on this level.

1. The student is asked to complete the sentence.

   Example:

   An orange is something to _____.

   ◯ see
   (*) taste
   ◯ feel

2. The student is given three sentences. She must decide which sentence tells about something that could happen.

   Example:

   ◯ The cat walked across the street.
   ◯ The boy spread his wings.
   (*) The bird sang in the tree.

Levels 13-14

This format is also included on these two levels.

3. The student is asked to find the word that is used in a similar way as the underlined word(s) in the given sentence.

Example:

The boy is in a fog.

a. lost

*b. confused

c. wet

d. cold

Levels 13-19

This format is used on all levels.

4. The student is asked to read a passage carefully, and then answer the questions that follow. There will be several different kinds of passages.

Example:

When the children woke up, they were excited because it had snowed during the night. They looked forward to making a snowman after breakfast.

When did it snow?

f. last week

g. yesterday

*h. during the night

j. tomorrow

Spelling

Level 13

The following format is used on this level only.

1. The student must decide if the underlined word in the sentence is right or wrong.

   Example:

   The dog barked. Right (*) Wrong ( )

Levels 14-19

This format is used for all other levels.

2. The student is given a complete sentence with two or three of the words underlined. The student must decide if one of these underlined words is misspelled. If not, he must choose the word "none."

   Example: The banker washed his durty car. None

   a *b c

Language Mechanics

Levels 13-19

The following two formats are used on all levels.

1. The student must indicate the section that contains a word to be capitalized or choose the word "none."

   Example:

   Billy told/ his friend / to meet him / at gino's. None

   a b c *d e

2. The student must decide which punctuation mark, if any, is missing.

   Example:

   What time is it

   , ? " None

   a b *c d e

Language Expression

Levels 13-19

The following two formats are used on all levels.

1. The student is asked to choose a word or words that best complete a sentence.

   Example: Give the test to _____ children.

   a. them

   *b. those

2. The student is asked to choose a part of a sentence that is the subject, and the part of the sentence that is the verb.

   Example: subject

   She saw the stars last night.

   *a b c d

   Example: verb

   She saw the stars last night.

   a *b c d

Levels 14-16

This format is also included on the above levels.

3. The student is asked to recognize a complete sentence, an incomplete sentence or a run-on sentence.

Example:

If the telephone does not ring.

a. run-on sentence

b. complete sentence

*c. not a complete sentence

Levels 14-19

This format is also included on the above levels.

4. The student is given a list of sentences, and must decide the best order of the 4 sentences which make up a paragraph.

Example:

1. Then he went to school.
2. Tim woke up at 6 o'clock.
3. He got out of bed.
4. He got dressed and ate breakfast.

a. 1, 2, 3, 4

*b. 2, 3, 4, 1

c. 2, 4, 3, 1

d. 4, 3, 2, 1

Levels 16-19

This format is also included on the above-levels.

5. The student is given a sentence that is called the topic sentence. He/She must choose the pair of sentences that develops the topic of the given sentence best.

Example: The play was a success for the students in our school.

*a. The students worked very hard at all rehearsals. They performed well during the performance.

b. This was the second time the school had done "Romeo and Juliet." It was a favorite.

c. The school sponsors many extracurricular events. The play is one of them.

d. The play was performed at night. It lasted two hours.

Levels 17-19

These two formats are also included on the above levels.

6. The student is supplied with three sentences. The student must choose the one that is most clearly expressed.

Example:

a. The teacher helped the boy to sit down with the broken arm.

*b. The teacher helped the boy with the broken arm to sit down.

c. The teacher helped to sit down the boy with the broken arm.

7. The student must choose a concluding statement from 4 alternatives, after reading the given paragraph.

Example:

Traditionally people bought their necessary purchases from merchants in various parts of the cities. With the development of the suburban shopping malls, people were able to find everything in one area.

*a. The shopping malls are hurting businesses in the cities.

b. Many shopping malls also have several movie theaters.

c. Many people live in the cities.

d. Shopping malls are usually close to large housing areas.

Mathematics Computation

Levels 13-19

There are four formats included on all levels.

The student is given exercises of addition, subtraction, multiplication and division. Each operation is presented in two ways, horizontally and vertically, and "none of the above" is always an alternative.

1. Addition

Example: 326 + 2 =

a. 324
b. 163
c. 652
*d. 328
e. None of the above

Example: 224
+ 4

a. 220
b. 56
c. 896
*d. 228
e. None of the above

2. Subtraction

Example: 404 - 400 =

a. 100
b. 804
*c. 4
d. 8
e. None of the above

Example: 320
- 300

a. 100
b. 620
*c. 20
d. 40
e. None of the above

3. Multiplication

Example: 50 x 2 =

a. 25
*b. 100
c. 48
d. 52
e. None of the above

Example: 16
x4

a. 4
b. 164
c. 12
d. 20
*e. None of the above

4. Division

Example: $48 \div 4 =$

*a. 12

b. 24

c. 44

d. 52

e. None of the above

Example: $2\overline{)10} =$

*a. 5

b. 20

c. 8

d. 12

e. None of the above

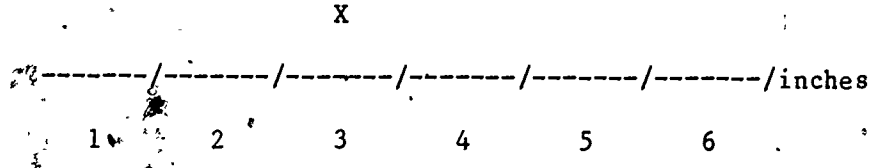
Mathematics Concepts and Application

Levels 13-19

One format is used on the subtest, Mathematics Concepts and Application, on all levels.

1. The student is asked to choose the correct answer to a question which sometimes uses a picture, table or diagram to illustrate the problem. The answers do not always require a numerical response.

Example:



How many inches are indicated by the X?

a. 6

b. 1

*c. 3

d. 4

e. None of the above.

Reference Skills

Levels 14-19

One format is used on the subtest, Reference Skills, on the above levels. There is no Reference Skills subtest on Level 13.

1. The student is given a sample reference material and then asked to answer questions that relate to it. Several different samples are used throughout the test. The samples include maps, indexes, tables, dictionary pages, library catalog cards, diagrams, forms, title pages and reader's guides.

Example:

GROCERY ITEMS	COST
eggs/1b.	$ .87
milk/half gal.	.92
bread/loaf	.75
cheese/1b.	.99

How much does a loaf of bread cost?

a. $ .87

b. .92

*c. .75

d. .99

CHAPTER 2

REPORTING TEST RESULTS

This chapter contains information about reports on test performance that are provided to schools. Additional test information is reported each year in the Annual Test Report. The discussion of each report covers three areas--the questions that can be answered by the report, an explanation of the data reported, and a description of how to use the report. Technical terms used here are explained in Chapter 3.

2A. SCHOOL FREQUENCY DISTRIBUTION AND MEAN SCORE REPORT

These reports can be used to answer the following questions.

1. What is the average or typical test score for a school?

2. Does the school have subject areas in which it is performing especially well or especially poorly?

3. How much variation is there in the test scores for a school?

Data Reported

The School Frequency Distribution contains the number (frequency) of students attaining each raw score on each subtest and total. Table 2.1 is a sample report that illustrates the parts of the report that will always be provided. Presented with each raw score is its scale score, national percentile rank, national stanine and normal curve equivalent. MCPS percentile ranks and stanines may also be reported. These are computed each year and thus are not directly comparable from year to year

because they are based on different students each time. The final column on the report presents the cumulative frequency. This is the number of students in the school who scored at or below the raw score listed in that row.

At the bottom of the report of each subtest, the mean and standard deviation are presented for each score for which they can be computed. To the left of the raw score is an indication of where the median (MED) and first (Q1) and third (Q3) quartiles fall.

TABLE 2.1

SAMPLE FREQUENCY DISTRIBUTION

	Raw Score	Scale Score	National Stanine	National Rank Percentile	Normal Curve Equivalent	Frequency	Cumulative Frequency
	30	622	9	99	99	1	12
	28	589	8	92	80	1	11
Q3	25	553	7	86	73	2	10
	24	546	6	76	65	1	8
MED	22	530	5	57	54	3	7
Q1	20	517	4	35	42	2	4
	16	492	2	10	23	2	2
MEAN	22.50	539.25			56.83		
SD	4.07	35.89			21.7141		

The mean scores and their national percentile ranks are listed in the Mean Score Report that is published as part of the Annual School Progress Report and the countywide Annual Test Report. A sample of The Mean Score Report is shown in Table 2.2.[1]

[1]The mean scale score is shown in this example instead of the mean grade equivalent that was reported for the Iowa Tests of Basic Skills. The scale score is the recommended score to use for this purpose because it is on an equal-interval scale.

TABLE.2.2.

SCHOOL MEAN SCORES

	Grade 3		Grade 6	
	Scale Score (370)*	Percentile Rank (50)*	Scale Score (449)*	Percentile Rank (50)*
TOTAL BATTERY	387	68	473	69
Phonics Analysis	399	65		
Structural Analysis	405	71		
Reading Vocabulary	375	42	452	46
Reading Comprehension	416	63	498	68
TOTAL READING	389	61	470	60
Spelling	437	65	546	77
Language Mechanics	460	65	546	77
Language Expression	446	64	536	78
TOTAL LANGUAGE	440	66	536	79
Math Computation	348	70	451	65
Math Concepts and Applications	410	72	479	72
TOTAL MATH	381	72	465	69
Reference Skills			523	78

*Mean for the national norm group for the Total Battery

Uses of These Report

The typical score for a school can be the mean or median; generally they are equal or close to being equal. The typical score can be used to determine the strengths and weaknesses in each school's program. Percentile ranks (PR) of the typical scores should be compared to make this determination. However, PRs can only indicate which score is higher. The size of differences between PRs should not be used. If one wants to compare differences between subtests, (e.g., Is the difference

between Math Computation and Reading Comprehension larger than the difference between Language Mechanics and Spelling?) the mean or median computed using the normal curve equivalent (NCE) scale should be used.[2]

The most meaningful indicator of score variation in a school can be obtained by using the range of scores between the first (Q1) and third (Q3) quartiles, called the quartile range. This shows where the middle fifty percent of the scores in the school were. This range can indicate if most of the students in the school have a similar achievement level (a homogeneous school) or if the achievement levels are spread over a wide range (a heterogeneous school). The difference between Q1 and Q3 should be computed using NCE scores.

The School Mean Score Report brings together the typical school scores on all subtests so that the identification of strengths and weaknesses is apparent. As explained previously, this identification should be done by comparing the percentile ranks of each subtest. The scale scores cannot be used for this purpose because they do not indicate the same level of achievement across subtests. The results in Table 2.2 show that the school may want to take a close look at how they teach vocabulary because the percentile rank for that subtest is somewhat below the others.

---

[2]Percentile ranks are not on an equal interval scale and thus a 10 point difference does not have the same meaning at all points on the scale. NCEs are on an equal interval scale. See Chapter 3 for additional discussion.

Comparison of the mean scores across grades should be done with caution. Such a comparison provides descriptive information only and does not provide information about program effectiveness. This is because each grade group is made up of different students with different ability levels and backgrounds. The score differences can be caused by these factors and not be related to how well the students are taught. A better way to use test data to look at program effectiveness is explained in Section 2B which deals with longitudinal analysis.

2B. REPORTING DATA FROM LONGITUDINAL ANALYSES

The longitudinal analyses of school test data can be used to help answer the following questions:

1. Have students who have been in the same school for at least two years been able to maintain, or improve, their standing relative to the national norm?
2. How are school test scores affected by student transfers both in and out of the school in the years between test administrations?
3. Do the scores of the transferring groups indicate meaningful changes in the school's population?

The answer to the first question provides the best information from norm-referenced test data for looking at the effectiveness of a school program with regard to the objectives measured by the California Achievement Tests. Data relating to all three questions can be used to determine if programmatic changes are needed.

Data Reported

The results are reported as two different types of data--longitudinal (L) and nonlongitudinal (NL). In the results for School A (Table 2.3) the longitudinal data are the results from one group of students who were tested in the same school both years (i.e., for two consecutive test administrations). The nonlongitudinal data represents results from two

groups of students who were each tested in the school only one year. The group in the lower grade (3) transferred out of the school sometime after the first test administration. The group tested in the higher grade (5) transferred into the school sometime after the first test administration. Remember that the two nonlongitudinal groups are composed of completely different students.

TABLE 2.3

SCHOOL A

LONGITUDINAL ANALYSIS: GRADE 3 and 5: CALIFORNIA ACHIEVEMENT TEST

			Students Tested in This School Both Years			Students Tested in This School Only One Year		
	Grade	Year	Number Taking Test	NCE Mean	Percentile Rank of Mean	Number Taking Test	NCE Mean	Percentile Rank of Mean
TOTAL READING	3	1978	47	75	88	33	47	44
	5	1980	47	74	87	37	+69	82
TOTAL LANGUAGE	3	1978	46	72	85	33	49	49
	5	1980	46	+83	94	35	+65	76
TOTAL MATH	3	1978	47	76	89	35	53	55
	5	1980	47	78	91	38	+76	89
TOTAL BATTERY	3	1978	44	75	88	32	50	50
	5	1980	44	79	92	34	+70	83

Mean scores are presented for both the longitudinal and nonlongitudinal groups on the California Total Reading, Total Language, Total Mathematics, and Total Battery. The means are computed and reported using the Normal Curve Equivalent (NCE) scale.[3] Percentile ranks for the mean scores are also reported because they provide an easy to understand frame of reference. Also reported are the number of students in each group. The data for any group with fewer than 35 students should be viewed with caution. Data are not reported for groups of less than 10 because such results would be very unstable.

The rows in the tables separate grades; the columns separate the L group from the NL groups.

Use of Longitudinal Data

Analysis of longitudinal data can provide an indication of the effectiveness of a school's instructional program. Score trends within a school provide the best information when using these data. To determine the score trend for the L groups, NCE means should be compared for the

[3]The NCE scale is used because it is more appropriate for looking at score differences than are grade equivalent scores or percentile ranks. This is because on the NCE scale there is an equal interval between all values. This is not true of grade equivalent or percentile ranks. Chapter 3 has an extensive discussion of these terms.

two grades tested. The expected trend is that students will maintain the same NCE score or percentile rank, within error limits, from one grade to the other.[4] Substantial deviations from this expected pattern should be considered indications of possible strengths or weaknesses in the school program. Substantial is defined here as greater than 7 NCE points.[5]

The data shown in Table 2.3 for School A demonstrate how to interpret this report. The increase in Total Language is the only substantial change for the school. (This is indicated by a "+".) The mean score increased 11 NCE points, somewhat above the 7 point standard. This could indicate that the school has an especially strong program for teaching the punctuation, capitalization, and usage skills measured by the California. As a group, these students appear to be making satisfactory progress in the other areas of the California.

Any declines of more than 7 points should be considered indications of areas where the school may have to put special emphasis. These are indicated by a "-". The results for School B (Table 2.4) show such a decline for Total Reading.

---

[4]This expected trend is sometimes adjusted for reasons explained later in this section.

[5]Seven NCE points is one-third of a standard deviation. This standard is often used as an indication of educationally meaningful change for group data.

TABLE 2.4

SCHOOL B

LONGITUDINAL ANALYSIS: GRADE 7 and 9: CALIFORNIA ACHIEVEMENT TEST

			Students Tested in This School Both Years			Students Tested in This School Only One Year		
	Grade	Year	Number Taking Test	NCE Mean	Percentile Rank of Mean	Number Taking Test	NCE Mean	Percentile Rank of Mean
TOTAL READING	7	1978	253	62	71	28	63	73
	9	1980	253	-48	47	31	-32	20
TOTAL LANGUAGE	7	1978	251	59	67	27	66	78
	9	1980	251	54	58	31	-34	23
TOTAL MATH	7	1978	254	56	62	28	67	79
	9	1980	254	53	56	32	-29	16
TOTAL BATTERY	7	1978	248	57	63	28	64	75
	5	1980	248	51	52	31	-30	17

Additional insight into the meaning of longitudinal results can be obtained by comparing the trends for a school with those of schools that had similar starting points (e.g., Grade 7 in the School B example). Similar starting points could be defined as any scores within 7 NCE points of the score for the school being studied. This comparison can be helpful because the relationship between level of achievement and performance on a standardized test is not always the same at all levels of achievement.

For example, it is likely that groups with very high scores in Grade 7 may tend to show a decline over two or three years. Therefore, if a high achieving school shows a substantial decline it can be useful (probably comforting) to know if other high achieving schools show the same trend. At the same time, if the other schools do not show this trend, the need for improvement in the school with declining scores is emphasized.

Use of Nonlongitudinal Data

Nonlongitudinal data can be used to assess the effects of student transfers on school test scores. As with the longitudinal data, the trend of scores, not the absolute values, is the most useful information and a change of more than 7 NCE points should be considered a substantial change. Here, however, substantial changes most likely reflect population shifts, not the strength or weakness of an instructional program. Obviously, significant population shifts indicate that programs might have to be modified to meet the needs of the incoming students.

The results for students tested only one year in School A (Table 2.3) show a meaningful change in the school population. The score increase from third to fifth grade was considerably larger than 7 points in all cases. Additionally the group of students transferring into the school (i.e., those in Grade 5) represents a large portion (greater than 40 percent) of the fifth grade in that school. This means that the score trend will have an effect on the overall school averages. When the NL

and L results are compared, it can be seen that the change in population is toward more homogeneity in the class tested in Grade 5. The students who left the school sometime after third grade testing were scoring well below the stable group who remained in the school for both test administrations. However, those who transferred in after third grade testing scored almost as high as the stable group. Thus, the overall school average will increase and the achievement levels of the students in Grade 5 are more similar within that group than they were in Grade 3.

The nonlongitudinal results for School B also show substantial performance differences between students transferring out (Grade 7) and those transfefring in (Grade 9). In this case the trend is a decline. The overall school average will be affected only a small amount by this difference because the transfers make up only about 11 percent of the students tested in each grade. However, this does not mean that these results should be ignored. The transfer students tested in Grade 9 have achievement levels somewhat below the other students in the school and this difference has to be considered when planning instruction. These students may very well need an instructional program quite different from that which is appropriate for most of the students in the stable group.

Adjustment of Expected Trend

When reviewing longitudinal and nonlongitudinal data one additional factor should be considered--the overall county trends. The county longitudinal trend is the difference of scores between two test administrations for all students in the same school both times. The county nonlongitudinal trend is the difference of scores between two test administrations for all students tested only once in a school. This trend should be used to adjust the expected trend of equal NCE scores that was previously discussed. This is because that expectation of equal percentile ranks across grades could be affected by factors such as sampling error in the test norms and varying degrees of match between the test and curriculum at different grade levels. It is possible that these factors actually make the fifth grade test a little more difficult for MCPS students than the third grade test. If such factors are operating on MCPS test scores it would be unfair to hold to the equal NCE expectation. Therefore, the 7 point standard should be based on the difference from the county trend. For example, if the county trend is a 3 point decline, then a school score decline would not be considered substantial until it was greater than 10 points. On that same test a substantial increase would be anything greater than 4 points (i.e., more than 7 points above a 3 point decline).

2C. PERCENT CORRECT BY OBJECTIVE

The Percent Correct By Objectives Report provides information that can be used to answer the following question:

> Within each subtest, are there any skill areas (objectives) in which the school is performing especially well or especially poorly?

Data Reported

Six data elements are reported for each objective on each subtest. These are defined below.

Number of Items -- the number of questions measuring that objective according to the publisher's classification.

School Percent Correct -- the average percent of correct answers for the questions measuring that objective, for the school.

Area Percent Correct -- the average percent of correct answers for the questions measuring that objective, for the administrative area.

County Percent Correct -- the average percent of correct answers on the questions measuring that objective, for the county.

Norm Percent Correct -- the average percent of correct answers on the questions measuring that objective, for the national norming sample when the test was standardized. (This information is not available for some standardized tests.)

County/School Differences -- the result of subtracting the county percent correct from the school percent correct. If the school percent is higher the result will be positive, if it is lower the result will be negative.

Use of this report

The information needed to evaluate a school's performance on objectives is found in the last column of the report, County/School Differences shown in Table 2.5. It is necessary to use these particular data because, on norm-referenced tests, objectives are measured by questions of varying degrees of difficulty. One objective may have five easy questions; another objective may have five very difficult questions. This means that a school may have 30 percent more correct answers for one objective than another simply because it is measured by easier questions and not because that objective is being taught more effectively. In that type of situation the use of School Percent Correct to determine strengths and weaknesses will provide misleading information. To overcome this problem an estimate of the difficulty of the questions

measuring each objective must be determined. The County Percent Correct is used as this estimate. Thus, if the county average is low, the school's average can be expected to be low.

It is assumed that when the values in the County/School Difference column are approximately the same for all objectives, the school is teaching all of the objectives about equally well. If there is a substantial difference between the values in this column, then it is possible that areas of strength or weakness have been identified. Substantial difference is defined here as 20 percent.[6] If the difference between the highest and lowest value in this column is more than 20 percent, the test results may have identified an objective that is taught especially well or especially poorly.

Table 2.5 shows a report for the Language Mechanics section of the California Achievement Test. The school shown had a higher percent correct than the county on all five objectives. However, this does not necessarily mean that this school is performing well on all objectives. Actually, the performance on the first objective is considerably lower, when compared to the county performance, than on the other objectives. The school may want to look into how it teaches the skills measured by that objective.

---

[6]This represents a statistically significant difference if there are at least 50 students in the grade tested. If the group is smaller, one should allow a slightly larger difference.

Table 2.5

LANGUAGE MECHANICS

OBJECTIVE	NUMBER OF ITEMS	SCHOOL % CORRECT	AREA % CORRECT	COUNTY % CORRECT	NORM % CORRECT	COUNTY/ SCHOOL DIFFERENCES
I, Proper Nouns, Adjectives	6	79	79	77	75	+ 2
Beginning Words, Titles	6	68	65	45	51	+23
End Marks	7	80	60	56	69	+24
Comma	6	77	50	47	53	+30
Quotation Marks	3	92	65	68	56	+24

Table 2.6 presents the results for a school in which performance was below the county for all objectives. This does not necessarily mean the school's instructional program is poor. Students in this school could be scoring below the county average for various reasons that have little to do with the quality of the instructional program for the year. The knowledge and skills students bring to school are examples of factors over which the school has little control. The results for this school indicate that the comma objective is learned better than the other objectives on this subtest. The County/School Difference for that objective is 28 points higher than the low difference of -30 for the last objective--Quotation Marks.

Table 2.6

LANGUAGE MECHANICS

OBJECTIVE	NUMBER OF ITEMS	SCHOOL % CORRECT	AREA % CORRECT	COUNTY % CORRECT	NORM % CORRECT	COUNTY/ SCHOOL DIFFERENCES
I, Proper Nouns, Adjectives	6	52	79	77	75	-25
Beginning Words, Titles	6	22	65	45	51	-23
End Marks	7	30	60	56	69	-26
Comma	6	45	50	47	53	- 2
Quotation Marks	3	38	65	68	56	-30

Additional Considerations

The number of questions measuring an objective should be considered when using this report. The guideline often used is that, if an objective is measured by fewer than 5 items, the coverage of the objective is questionable. Results for any such objectives should be used with caution.

2D. INDIVIDUAL TEST REPORT

The Individual Test Report presents information that can be used to answer the following question.

> Does an individual student have any subject areas in which he/she shows particular strength or weakness?

Data Reported

The student's test performance on the subtests and combinations of subtests is presented in three ways. These are briefly explained below and in greater detail in Chapter 3.

> Stanines — The stanine range is divided into 9 units. Each stanine includes several possible test scores. This scale is often used to report results for individual students because it generally is not affected by small score variation caused by test error.[7]

[7] See Standard Error of Measurement in Chapter 3.

National Percentile Rank -- The percentile rank range is divided into 99 units. Percentile ranks indicate the percentage of students in a group who scored the same or less than the student whose scores are being reported. In the case of the California Achievement Tests, the reference group is the sample of students on whom the test was developed in 1977. A score of 65 indicates that the student did as well as or better than 65 percent of the students in the reference group.

Score Band -- Score bands represent a range of test scores around a student's score. They indicate the amount that a student's score may reasonably be expected to vary due to test error. If a student was feeling poorly the day of the test, his score may have fallen to the lower end of the band. If a student made a couple of lucky guesses, her score may have been at the upper end of the band.

Use of the Report

The question regarding an individual's strengths or weaknesses can be answered in two ways. The first is a comparison of the score bands for each subtest with the score band for the Total Battery. This provides information about the subject areas in which the student is strongest and weakest. The second way of answering this question uses the stanines (or percentile ranks) and provides an indication of strengths and weaknesses compared to the national norm group.

The rationale for comparing the subtest bands to the Total Battery band is based on the assumption that the Total Battery score represents the student's overall level of achievement and that any marked deviation from that overall level is noteworthy. The results of this comparison can be interpreted in the following ways.

1. A subtest score band which is completely below the band for the Total Battery indicates that the student appears to be doing poorer in learning the skills in that subtest than in learning other skills measured by the California Achievement Tests.

2. A subtest score band which is completely above the Total Battery score band indicates that the student appears to be doing better in learning the skills measured by that subtest than in learning the other skills measured by the California Achievement Tests.

3. A subtest score band that overlaps the Total Battery score band is an indication that the student's achievement in the skills measured by that subtest is about average for that student.

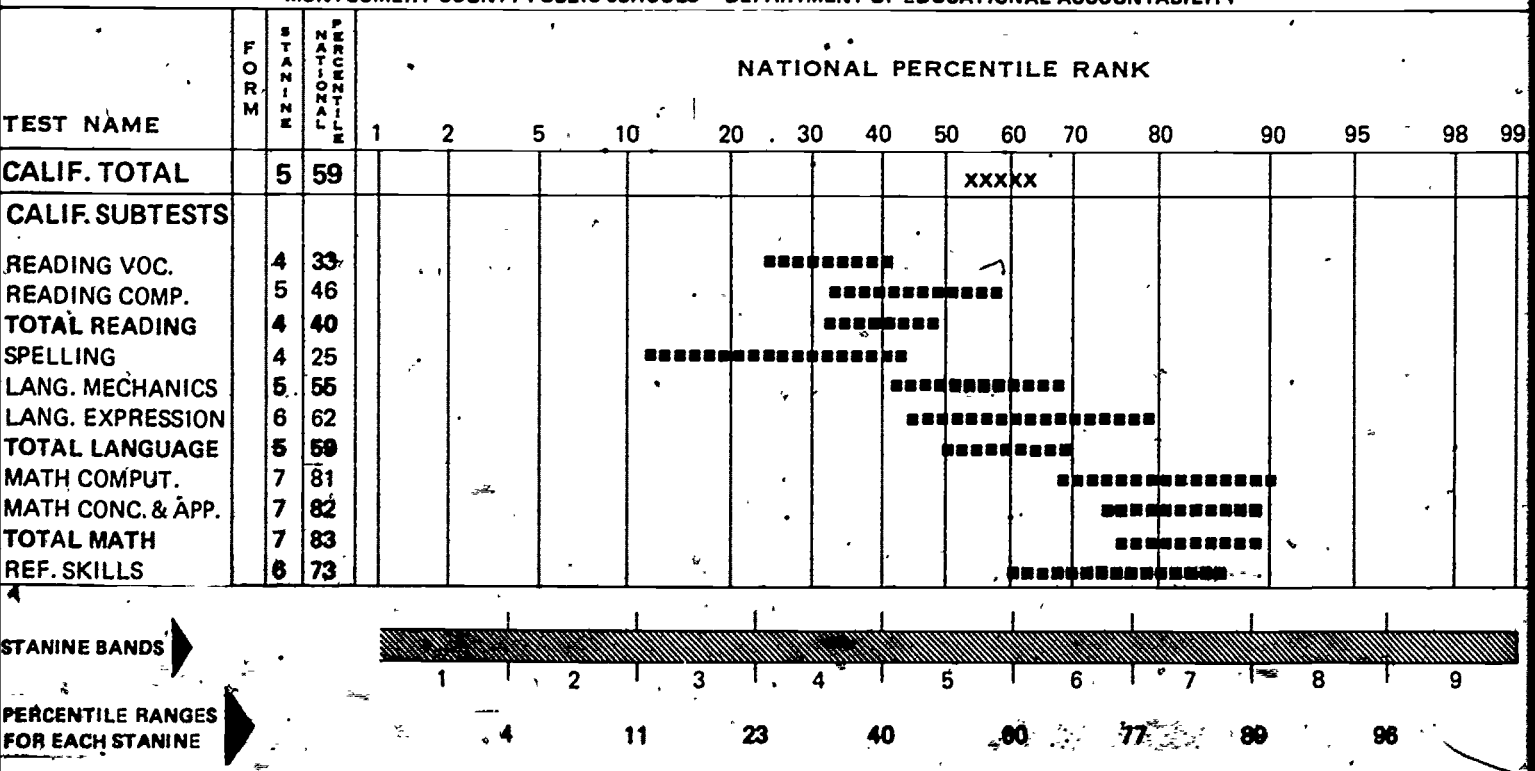
In Figure 2.1 , Reading Vocabulary is an area in which the student probably needs help, while Mathematics Concepts and Applications is an especially strong area.

High and low stanines provide indications of strengths and weaknesses regardless of the location of the score bands. Stanines of 3 or lower should be viewed as an indication that the student may be having trouble learning what is measured by the subtest. Stanines of 7 or higher mean the student is doing very well in learning the skills measured by that

FIGURE 2.1

INDIVIDUAL TEST REPORT

MONTGOMERY COUNTY PUBLIC SCHOOLS – DEPARTMENT OF EDUCATIONAL ACCOUNTABILITY



subtest. This is true even if the band for the subtest is completely below the Total Battery band. Such a pattern (i.e., subtest with a high stanine and score band below the Total Battery score band) can be interpreted as indicating two things -- (1) the student has high overall achievement and (2) his performance on this subtest is weaker than on the other subtests.

Use of Report With Parents

A copy of the "Individual Test Report" is sent to the parent or guardian for each student. Below are some possible questions parents may have upon review of this report.

There are at least two technical questions that parents are likely to ask with regard to the reporting of band scores. First, they may want to know why the bands for some subtests are larger than for others. This difference in width occurs because some subtests have larger components of measurement error. That is, they may contain more difficult questions that will cause students to guess more often. Guessing means students may get credit for knowing something they do not know. The length of the band can be shortened if the score is at the top or bottom of the percentile scale. This is because the student scored so high (or low) that even when the error factor is subtracted (or added) the band still does not extend beyond the 99th (or 1st) percentile.

Another question could come from parents who notice that for some subtests the student's actual score is not exactly in the middle of the band. This variation in score position occurs because percentile ranks are not equal distances apart. A more thorough explanation relating to this situation can be found in Chapter 3 in the discussion of percentile ranks.

2E. PROMINENT GUIDELINES FOR INTERPRETING TEST DATA

When reviewing test data, it may be helpful to employ the following guidelines. These may be particularly useful when interpreting test results to parents.

- Individual test scores are only estimates of student performance; the scores are subject to substantial measurement errors. This is why score bands are much more accurate ways of presenting test results than are numeric values.

- Norm-referenced test scores indicate a child's relative achievement or performance status, compared to students of similar age and grade level in the nation.

- Percentile ranks are derived by comparing a child's scores with those of students in the nation selected to establish the norms at some time in the past. The child's scores are not being compared to those of students currently in his or her grade.

- Norm-referenced tests provide an estimate of which students know the most about the content included in the test. These tests do not define in a very specific way what students know. Criterion-referenced tests are needed to serve the latter purpose.

- Standardized tests measure only some of the basic content skills common to curricula throughout the country. They do not measure the full curriculum presented in a particular class, school, or district.

- Percentile ranks do not refer to the percent of questions answered correctly, but to the percent of students whose performance an individual student has equaled or surpassed.

- Tests are not perfect. A percentile rank of 95 does not represent performance that is always superior to that represented by a percentile rank of 94 or 93.

- Test scores are not sufficiently precise to permit the ranking of students, except for very large differences (e.g., the fourth stanine vs. the seventh stanine).

- A stanine score of 7 or higher generally reflects strength in the area tested, while a stanine score of 3 or lower may indicate a potential problem.

- No one is expected to know everything on a norm-referenced test. Some items are purposely designed to be difficult.

- Avoid comparing a child's scores with those of his or her friends because of error in the scores and the confidentiality of the test scores.

CHAPTER 3

TECHNICAL TESTING TERMS

This chapter can serve as a reference for the technical testing terms used throughout this handbook and in other materials dealing with testing. The terms are defined; their uses are stated; and precautions about their interpretation are provided. 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 skills. 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).

Precaution(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:

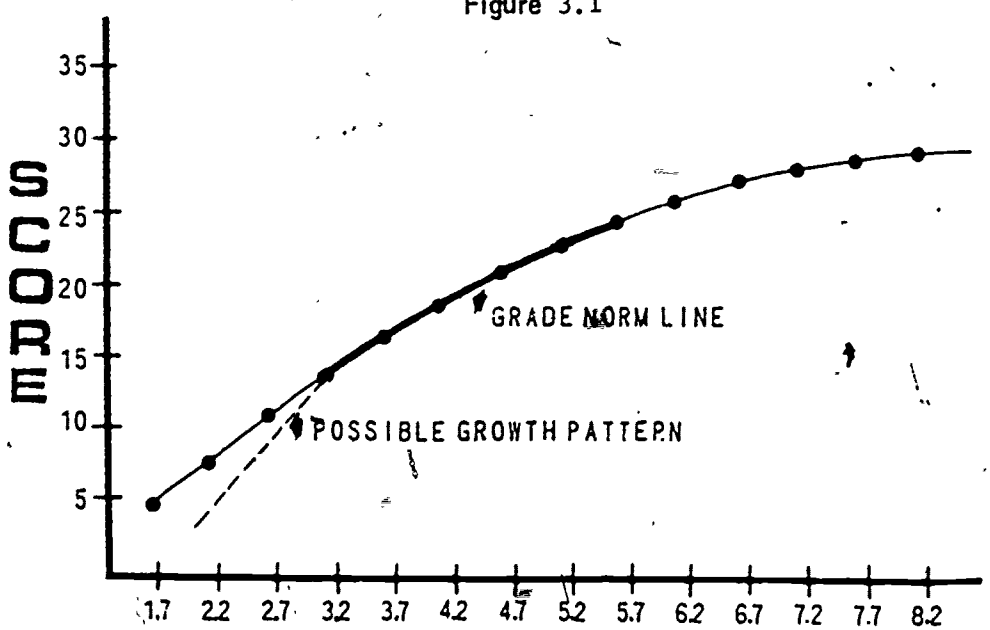
Number	Month
0	September
1	October
2	November
3	December
4	January
5	February
6	March
7	April
8	May
9	June-August

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 "." 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 3.1



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 ITBS seventh grade spelling test.

Grade	Test	Raw Score	Grade Equivalent	Difference in Grade Equiv.
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 that a student who has the same PR in Grades 3 and 5 will probably be further 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 as they have different meanings. For example, mathematics is more grade related than reading; and, therefore, the GEs are generally less spread out for math than reading.

7. Grade equivalents should not be compared across different tests because they may have different meanings due to 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 a measure of the performance of the "typical" student in a group.

Precaution

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. Half of the scores are above the median, half are below. It 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.

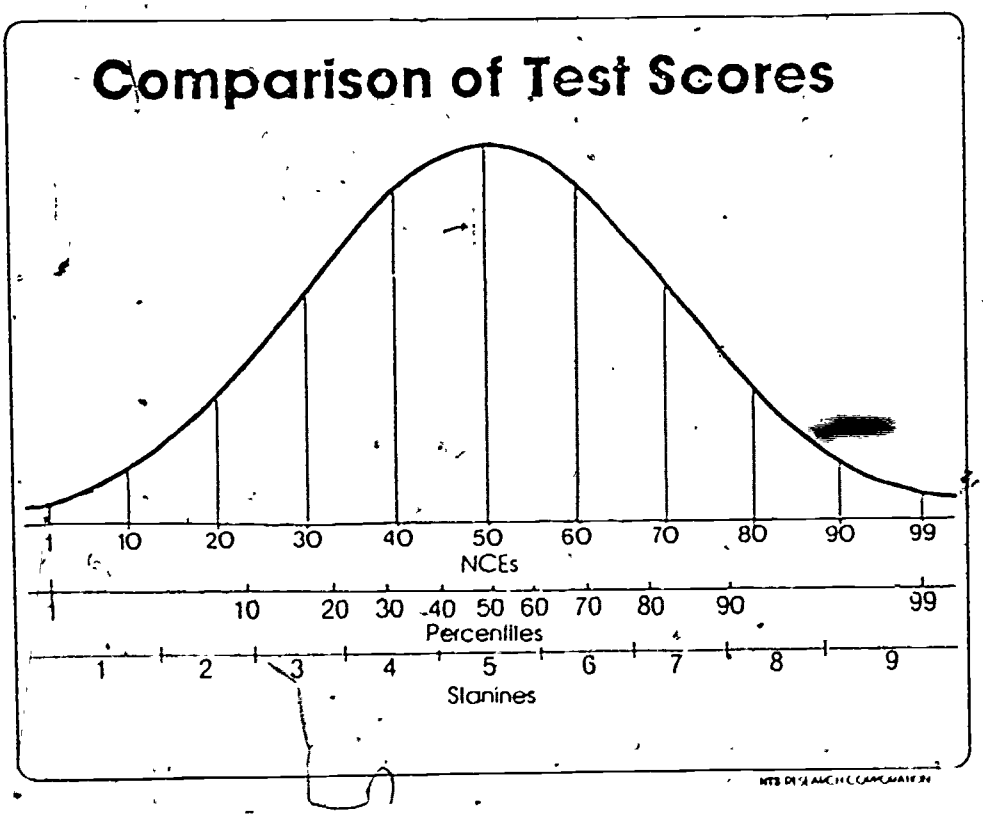
2. Use of the median provides no information about the spread of scores.

NORMAL CURVE

Definition

A normal curve is a distribution of scores or values which, in graphic form, is bell-shaped as shown in Figure 3.2. In a normal curve distribution, the mean and the median are at the same point. The majority of the scores are clustered around the mean/median. Sixty-eight percent of the scores are within one standard deviation of the mean/median, and 95 percent are within two standard deviations. Scores which are more than three standard deviations from the mean/median are rather rare, occurring less than 1 percent of the time.

Figure 3.2



Uses

Because of its well-documented statistical properties, the normal curve distribution is often used in reporting test scores as an aid in interpreting scores of groups or individuals.

Precautions

The normal curve distribution is a statistical or mathematical ideal. It is not a graphic description of what a particular distribution should be; distributions which do not conform to the normal curve are not "abnormal." Many variables can affect the distribution of a particular set of scores: test content, difficulty of the test items, suitability of the test for the group to which it is administered.

NORMAL CURVE EQUIVALENT SCORES (NCE)

Definition

NCEs divide the normal distribution into 99 segments, units, or scores (Figure 4.2). 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 4.2.

Uses

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

A test 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.

---

[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.

Use

Norm-referenced tests provide information about which students know the most about the content included on the test.

Precaution(s)

1. A good NRT is designed to enable between 40-70 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 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.

NORMS

Definition

Statistics that describe the test performance of specified groups, such as students in a given grade, age range, type of community, etc.

Use

Norms provide a way of relating raw scores to a more meaningful score scale, such as percentile ranks, stanines, grade equivalents, or a standard score, so that it can be determined how a student performed relative to a "representative" sample of students similar in some way.

Precaution(s)

1. Norming samples cannot be perfectly representative of a large group of students. For most major standardized tests, publishers use sophisticated sampling procedures to determine the norming sample. However, there will always be a small error factor. This means that caution must be used when comparing the scores from two different tests or even from two levels of the same test because the levels may not have used the same group of students. The following is an example of what might happen because of this. If the students in the norming sample for Test A are brighter than those in the sample for Test B, the norms for the two tests will not be equivalent. A student who then takes both tests will be likely to attain a lower percentile rank on Test A because he/she is being compared to a brighter group of students on a test which has "more difficult" norms.

2. Test publishers often provide norms for different times of the year such as fall, winter, and spring. However, they may not have used a norming sample at all of these times, which means that some of the norms are estimates. A test manual should be consulted to determine when a given test was normed. Estimated norms for any other time of year should be viewed with caution.

3. Test norms are not necessarily derived every year, and therefore some norms may be several years old. However, it is common practice to compare current student performance on a given test with the performance of the national norming sample. Caution must therefore be exercised in interpreting the meaning of an individual's status. For example, a student who took a test in 1978 and who achieved a percentile rank of 60 probably did not score higher than 60 percent of the students taking the test in 1978. Rather, the individual scored higher than 60 percent of the students in the norming sample who took the test in the past, for example in 1970.

4. The above considerations may weaken the usefulness of older norms. If changes have occurred in curricula, current students may be better prepared in some skills or subjects than were students in the norming sample, less well prepared, or simply differently prepared. Thus, comparisons of percentile ranks across years may be clouded by changing curricula.

5. Norms are derived so that half of the representative group is expected to be below average. This means that half of the group will be below grade level, below a percentile rank of 50 and below the mean. Therefore, it is extremely difficult to have all of the students in any large group perform above the average.

PERCENTILE RANK (PR)

Definition

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 to the performance of students in the norming sample.

Precaution(s)

1. PRs should not be added or 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 is 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

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 (ie., addition, etc.) is a questionable procedure. Generally raw scores do not have the equal interval 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.

RELIABILITY

Definition

Reliability refers to the extent to which a test is consistent in what it measures. There are three major types of reliability, all expressed as a coefficient ranging from 0 (complete lack of consistency) to 1 (perfect consistency).

1. Internal consistency is the degree to which all the questions on a test measure the same thing. For example, a mathematics test that measures only addition of fractions will probably have a higher internal consistency coefficient than one that measures several different mathematical operations. This would be especially important for achievement tests that measure specific skills.

2. Stability is the degree to which a person will achieve the same score on a test that is taken twice within a time period of anything from a few days to a year or two. This is important in an instrument which measures a trait like natural ability which is not expected to change over time.

3. Equivalence is the degree to which a person will achieve the same score on two forms of the same test. This is important for any test in which two forms are to be used interchangeably.

Use

Reliability is a measure of the quality of a test.

Precaution(s)

The type of reliability appropriate for a given testing situation should be used.

SCALE SCORE (SS)

Definition

Scale Scores range from 0 to 999 and provide a link between all levels of the California Achievement Tests.

Uses

1. Scale scores can be subjected to arithmetic operations like Normal Curve Equivalent scores. Therefore, means can be computed and differences in SSs can be compared meaningfully.

2. Scale scores provide a way of comparing scores on different levels of the California Achievement Tests and, therefore, provide a way of measuring growth.

3. The capability of comparing results from different test levels also means that scale scores help to make out-of-level testing possible. This testing procedure allows for a student to take a test for a grade other than his own and still have results (percentile ranks and stanines) based on norms for his grade.

Precaution

1. Scale scores should not be used to compare scores in different subject areas. They were not developed so that equivalent scores in two subject areas would indicate equivalent levels of achievement. Any comparison of scale scores should be done within subject areas.

2. There are not "typical" scale scores for each grade or test level. In fact, the ranges of SSs in the various levels overlap considerably.

SIGN TEST

Definition

A test of statistical significance which is based on the number of increases (+) and decreases (-) in a set of comparisons. If the pattern of pluses and minuses deviates substantially from an even split, the pattern is considered significant.

Use

To determine if a pattern of increases and decreases deviates from an even split enough to indicate a significant trend.

Precaution

1. The sign test indicates only if the overall trend of increases and decreases is significant. It does not provide any information as to whether individual increases or decreases are significant.

2. The size of a difference is irrelevant. For example, this test does not differentiate between an increase of 1 point or 30 points. They both simply count as a plus.

STANDARD DEVIATION (SD)

Definition

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.

Precaution(s)

None

STANDARD ERROR OF MEASUREMENT (SEM)

Definition

The SEM is an estimate of the magnitude of error in a test score. Possible causes of error in scores include lucky or unlucky guesses, a student's not feeling well or failing to follow directions, the fact that test questions may be only a sample of those that could be asked, sloppiness, laziness, etc.

Use

1. The SEM provides a way of determining the possible fluctuation in test scores which would be obtained if an individual were to take the same test a number of times. It indicates how far a particular obtained score might deviate from the individual's "true" score (the score the individual would obtain if there were no error in the test). It is usually assumed that the scores obtained from repeated testing would conform to the normal curve distribution. Therefore, in practice, it is assumed that there is a probability of 68:100 that the "true" score is within one SEM of the obtained score and that there is a probability of 95:100 that the obtained score is within two SEMs of the obtained score.

2. The SEM can be used in significance testing to provide a way of determining whether differences in test scores or group mean scores are statistically significant (that they vary more than can be reasonably attributed to testing error).

Precaution(s)

None

STANINE

Definition

A stanine is one of the scores of a 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 3.2, each stanine has a range of corresponding percentile ranks or raw scores.

Use

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 3.2, 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.

STATISTICAL SIGNIFICANCE TEST

Definition

A significance test is a statistical procedure used to determine if two (or more) groups differ on a trait more than could normally be expected if testing error or sampling error were assumed to be the cause of the difference.

Use

Under highly controlled conditions (as in experiments, etc.), tests of statistical significance are used to test hypotheses. When variables cannot be controlled (as in the countywide testing program), the results from such a test are open to question.

Precaution(s)

1. Results of significance tests are reported as probability statements. If the reported probability is less than .01, the chance is less than 1:100 that the difference between groups can be attributed to testing error. If the probability is .001, the chance is less than 1:1000 that the difference can be attributed to testing error. However, there is always <u>some</u> chance (1:1000, etc.) that the difference was caused by error.

2. When a large number of tests of significance are performed, some differences will turn out to be statistically significant by chance alone. That is, since there is always some chance that a difference can be caused by error (1:20, 1:100, 1:1000, etc.), a certain number of significant differences can be expected to occur because of error. There is no way to determine if a particular statistically significant difference was or was not caused by error. Again, only a probability can be determined.

3. When tests of significance are used to evaluate the difference of means, the larger the group the smaller the difference in means needs to be for statistical significance. The smaller the group the larger the difference must be. For example, a difference of only 1-2 months on the grade equivalent scale, or a fraction of a raw-score point, will be statistically significant for groups of several thousand students. In contrast, a difference of as much as six months may be required for significance with a group of one hundred students. Because many of the comparisons in this report involve very large groups, no significant tests of differences and means were performed. While small differences would have been statistically significant, they would not have been educationally meaningful.

VALIDITY

Definition

The extent to which a test does the job for which it is used. There are three major types of validity that a test may possess.

1. Content validity is most important for achievement tests. This requires a test to contain questions that adequately reflect the content the test is supposed to measure.

2. Criterion-related validity is most important for placement tests, college admissions tests, or tests on which employment decisions are based. Performance on the test must be highly correlated with performance in the program, success in college, or success on the job for which the test is a screening instrument.

3. Construct validity is most important in psychological instruments. Tests of ability are examples of such instruments. Construct validity requires that the test adequately discriminate between people who do or do not have a particular trait.

Use

Validity is a measure or concept that helps one evaluate the quality of a test.

Precaution(s)

The type of validity appropriate for a given testing situation should be used.