ED 260 085 TM 850 280

AUTHOR	Burns, Marilyn
TITLE	What Do Test Scores Really Mean? Revised Issue. Publication No. 443.
INSTITUTION	Los Angeles Unified School District, Calif. Research and Evaluation Branch.
PUB DATE	Sep 84
NOTE	42p.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Achievement Tests; *Educational Testing; Elementary Secondary Education; *Scores; Scoring; *Standardized Tests; Testing Problems; *Test Interpretation; Test Norms; Test Reliability; Test Results; *Test Use; Test Validity

ABSTRACT

This guide--written for school administrators, teachers, counselors, parents, and the community--describes educational tests and measurements and test-related statistics. While specifically intended to help readers interpret the tests administered in the Los Angeles (California) Unified School District, this information may also be used to interpret other test results. Topics include: (1) why schools administer tests; (2) what tests measure; (3) objective tests and subjective tests; (4) standardized, norm-referenced, and critereon referenced achievement tests; (5) test reliability and validity; and (6) standard error of measurement. The section on scores describes raw scores, percent correct scores, percentile scores, scale scores, grade equivalent scores, stanines, and criterion referenced scores. Scores for groups (of students are also discussed: central tendency (mean, median, and mode scores) and variability (range, quartile rank and standard deviation). In the concluding section, an analysis of the real meaning of test scores is offered in which potential problems resulting from test use are cited and discussed. A 66-item glossary is appended. (GDC)

BEST COPY AVAILABLE

WHAT DO TEST SCORES REALLY MEAN?

PUBLICATION NO. 443

RESEARCH & EVALUATION BRANCH

LOS ANGELES UNIFIED SCHOOL DISTRICT

WHAT DO TEST SCORES REALLY MEAN?

PUBLICATION NO. 443

A Report Prepared by the

Research and Evaluation Branch

of the

LOS ANGELES UNIFIED SCHOOL DISTRICT

Original Issue: September 1974
Revised Issue: September 1976
Revised Issue: September 1984

LOS ANGELES UNIFIED SCHOOL DISTRICT

HARRY HANDLER
Superintendent

This report was prepared by
Marilyn Burns, Assistant Director
Testing Unit

APPROVED:

FLORALINE I. STEVENS
Director
Research and Evaluation Branch

JOSEPH P. LINSCOMB
Associate Superintendent, Instruction

TABLE OF CONTENTS

FOREWORD

"What Do Test Scores Really Mean" was revised and reissued because many school administrators, teachers, counselors, parents, and community persons need and use test information. In particular, these persons want to learn about test results as they relate to the tests that are administered to students in our school district. This publication is a mini-primer of information about tests and measurements and test-related statistics.

All teachers and administrators in this school district will receive a copy of this publication. In addition, this publication will be the basis for a television series for staff to upgrade their knowledge and use of tests.

I am especially appreciative of the talented persons responsible for this revised issue. Mrs. Marion Beller, retired Specialist, initiated the project, and Dr. Marilyn Burns, Assistant Director, Testing Unit, had the primary responsibility of writing and producing the final document. I am grateful to all persons in the Testing Unit who reviewed and then offered suggestions to improve the document.

I anticipate that the information in this publication will prove to be useful and needed throughout the ensuing years.

Floraline I. Stevens, Director
Research and Evaluation Branch

WHAT DO TEST SCORES REALLY MEAN?

1. WHY SHOUID WE KNOW ABOUT TESTS?

Tests and test scores are used in nearly all of today's schools. Despite occasional controversy about the meaning of test scores or the number of tests given, teachers, parents, and the public expect students to be tested and expect the test scores to provide information about individual students and about the quality of publicly supported schools and school districts.

This publication is designed to help readers interpret the tests administered in the schools of the Los Angeles Unified School District (LAUSD). It may also be used, however, to help readers interpret test scores reported in the media. By asking appropriate questions and looking for necessary details, each reader can become a wise "test consumer," knowledgeable about what information tests do and do not provide.

2. WHY DO SCHOOLS TEST?

Tests are important in evaluating individual student progress and local school and overall district achievement levels. It is necessary to have this kind of information for long-range instructional planning to meet educational needs of all children in the district. School administrators, teachers, and parents use test results for initiating program review and improvement, for compiling schools' needs assessments, and for providing one of the bases for instructional activities.

Some test scores are also used as one component in counseling and individual program planning. The scores are used by the district to assess how students in the LAUSD com are with their peers statewide and nationally and in decision making ial programs.

3. WHAT DO TESTS MEASURE?

In education, a test is a device or an instrument for measuring specific characteristics of individuals or groups by requiring the performance of a specified task or series of tasks. Tests are usually designed to measure one of the five following characteristics:

Achievement

An achievement test provides measurements of accomplishment in school subjects as shown by knowledge of important facts or ideas taught in the course. It may also measure a student's understanding of and ability to apply those facts and ideas.

Scholastic Aptitude, Mental Ability, or Intelligence

These three terms are used interchangeably, but scholastic aptitude seems to be the preferred usage. These tests attempt to identify and measure those kinds of capacities or qualities which may predict academic success.

Interest

A test which measures student interests does not involve right or wrong answers. What is often referred to as an interest inventory measures a student's likes and dislikes, or preferences. Instead of a score, the results are usually expressed as a profile of the relative strength of the stud(·‧ᐩ's var'ous academic or occupational interests. For many interest inventories there are norms which allow a comparison of the student's pattern of interests with those of his or her age and grade peers.

Attit.ıdes

A student's attitudes may greatly influence how he approaches problems both in and out of school. Attitude tests may give the teacher information on how a student feels about school, peers, self, and learning - usually on one or more scales since there are no right or wrong answers. Some attitude tests have norms which allow a comparison of the student's attitudes with those of age or grade peers.

'ersonality

Like interest inventories, there are no right or wrong answers on a personality test. The results are usually stated on a scale or series of scales. Since the use of personality scales and scholastic aptitude tests is limited to school psychologists and since the interest test is largely limited to local high school use, this publication will focus its attention on achievement tests.

4. WHAT IS AN OBJECTIVE TEST?

Labeling a test as objective or subjective refers to its physical format and to the way in which it is scored.

Objective Tests

Objective tests are the most frequently used tests today. An objective test consists of many questions, each of which has a predetermined correct answer that is to be selected by the test-taker from among two or more suggested answers. Multiple-choice tests and true-false tests are of this type. In a well-made objective test, all of the answer options must be plausible and equally well-written -- a time-consuming task.

An advantage of objective tests is that they may be scored quickly and accurately. They also lend themselves to machine scoring, which is why nearly all large scale testing programs use multiple choice objective tests. Machine scoring also allows for computer analysis of the data and the production of summaries which include all school, region, or district students.

<u>Subjective Tests</u>

The inference of the word "subjective" is that no standards for scoring can be used other than "teacher judgment." This is <u>not</u> true Subjective tests include those which require the student to construct an answer in response to the question or to write an essay on a given subject.

Subjective tests must be hand-scored, but a well-designed test will have scoring criteria prepared in advance, which make it more likely that scorer bias is minimized and that each student's response is scored fairly. The essay tests which are part of the district's WRITE:Sr competency test are scored in this objective way although they are, basically, subjective tests.

Many tests prepared by classroom teachers are subjective and have a useful function. Requiring the student to write an essay or paragraph taps skills which cannot be fully measured with an objective test. In some classes, a work sample, art project, or research paper is the best way to measure what has been learned.

5. WHAT IS A STANDARDIZED TEST?

Standard conditions of administration, standard timing, standard test questions, and standard answer documents are all part of the test situation which makes it possible to compare the test scores of a student in Los Angeles with those of other students in the standardization sample who have known characteristics such as age or grade. Standardization allows us to judge "how we are doing" against a known standard.

One type of standardized test is a norm-referenced test. The standard is a table of national norms developed by testing a large sample of students who are representative of all American students in the grade levels tested. This sample represents all student ability levels, all geographical areas, and all socioeconomic and racial/ethnic groups, proportionally, in the nation. Each test's manual describes the size and characteristics of the sample used to develop its norms.

The norms are developed by giving the sample of students a very large number of experimental questions which reflect curricula across the nation and by later reducing the questions to the number which appear in the published test. Careful analysis of each test item and selection of each test item by test development experts result in a test which permits comparison of other individuals and groups with a large sample of American students upon which the national norms have been based.

Criterion-referenced tests, such as the district's Survey of Essential Skills (SES), may also have standardized administration and scoring procedures even though an individual's performance is not compared with that of a norm group.

The largest number of tests administered in any school are non-standardized teacher-made tests which the teacher uses to assess instructional progress. In some cases, especially at the high school level where specialized subjects are taken separately, the teacher-made test may be replaced by a standarized, subject-matter test which can help assure the teacher that the students are achieving in geometry or first year chemistry, for instance, at a level recognized nationally as appropriate for that subject. Sometimes these tests are included with instructional materials for the subject, and at other times they must be separately purchased.

6. WHAT ARE THE DIFFERENT TYPES OF STANDARDIZED ACHIEVEMENT TESTS?

There are two major types of achievement tests in common use, norm-referenced and criterion-referenced. Each has its advantages and disadvantages. Each can provide valuable information if it is well designed and properly used.

Norm-Referenced Achievement Tests

This is the type of test given by almost all school districts in order to evaluate pupils and schools on a national norm scale. In Los Angeles, the Comprehensive Tests of Basic Skills (CTBS) has served this purpose for several years. An over-all district score can also be produced which makes it possible to compare it with other school districts. There is often great public curiosity about these comparisons.

Norm-referenced tests, such as CTBS, are designed for a particular grade level or levels and include many test items which reflect the instructional content for that grade level. They also contain test items which reflect the instructional content below and above the grade being tested. In this way, the achievement level within the grade level can be ranked from low (well below grade level) to high (well above grade level) and can illuminate small differences in achievement between students.

Some norm-referenced tests allow a comparison of schools and districts with a norm but do not provide for comparison of individual students. The California Assessment Program (CAP) tests are of this type. This is because the CAP test is a very long test which has been divided into 30 to 40 short tests. Since no individual student takes the entire test, only school, region, and district average scores can be generated for the California Assessment Program. The norm group for the CAP consists of al. 'alifornia students tested at that grade level.

Criterion-Referenced Achievement Tests

A test designed to show whether a student has learned a specific skill without comparison to other students is called a criterion-referenced test, or an objective-referenced test. Criterion-referenced tests may be subjective, objective, or a combination of both. They are designed to determine whether an individual student has met a specific instructional objective or criterion of performance. They do not show whether success at the task is typical for a particular grade level or how the performance of an individual compares with the performance of other individuals or groups of individuals.

The individual's progress is generally reported by listing the skills the individual has mastered, the mastery level of achievement having been previously identified.

In the Los Angeles Unified School District, state-mandated competency tests are of the criterion-referenced type. The tests used at the elementary level (grades 1-6) at each district school are the Survey of Essential Skills (SES). The SES measures the essential, basic skills required at each grade level. It is expected that all students will master the skills tested and score very high on the SES. Individual student reports list the district's curriculum continuum skills and the student's level of mastery in each skill.

For several years the district's criterion-referenced tests at the junior high level have been Performance Assessment in Reading (PAIR), Assessment of Skills in Computation (ASC), and Test of Performance in Composing/Enabling Skills (WRITE:Jr). Alternate junior high competency tests are being introduced which also require students to apply their mastery of the skills of reading, computing, and writing. Test results are reported in "ranges" of mastery.

At the senior high level, the criterion-referenced tests are Senior High Assessment of Reading Performance (SHARP), Test of Performance in Computational Skills (TOPICS), and Test of Performance in Composing/Enabling Skills (WRITE:Sr). These tests present simulated real life situations in which students apply their mastery of the skills of reading, computing, and writing. Test scores for senior high are "pass" or "fail" to meet the established criterion. Students have several opportunities to take and "pass" any test which they had previously failed.

7. WHAT DO THE DIFFERENT TYPES OF TEST SCORES MEAN?

Educators frequently refer to tests as measuring instruments because they use them to measure aptitude, achievement, or some other characteristic much as a thermometer measures temperature. Test scores serve the same purpose as the degree numbers and graduation lines of the thermometer. They show (within predictable limitations) the degree to which the characteristic being measured is present.

It is important to differentiate between individual student test scores and the group scores reported for class, school, or school district. Each type has its own precise terminology which must be understood before the meaning of the score is revealed.

Scores for Individual Students

There are many ways of reporting how well students do on tests. Each type of score is a different way of looking at the test performance being measured, and each gives a different shade of meaning. The most commonly encountered types of student scores and their common abbreviations are:

raw score (RS)

percent correct (%)

percentile scores (%ile)

scale scores (SS)

grade equivalent scores (GE)

stanines (STA 9)

criterion-referenced scores

Raw Scores

The number of correct answers a student gets on a test is called the "raw score." In some tests different items are "weighted" so that difficult items are assigned more points than easy items. The raw score, then, is the total number of points earned. In a timed test, such as in typing, the raw score may include both the time elapsed and the number of errors.

Raw scores are often used by teachers who tally the scores of their students on each item or group of items. This type of "item analysis" aids the teacher in planning instruction, both ongoing and remedial, in the academic area tested.

Percent Correct Scores

The interpretation of a raw score is impossible without knowing the total number of questions on the test. A raw score of 25 correct answers out of 25 questions has a different meaning than 25 out of 50. The difference can be stated as the percent correct. If the raw score is 25 out of 25, the percent correct would be 100% -- the highest possible score. However, if the raw score is 25 out of 50, the percent correct would be 50%.

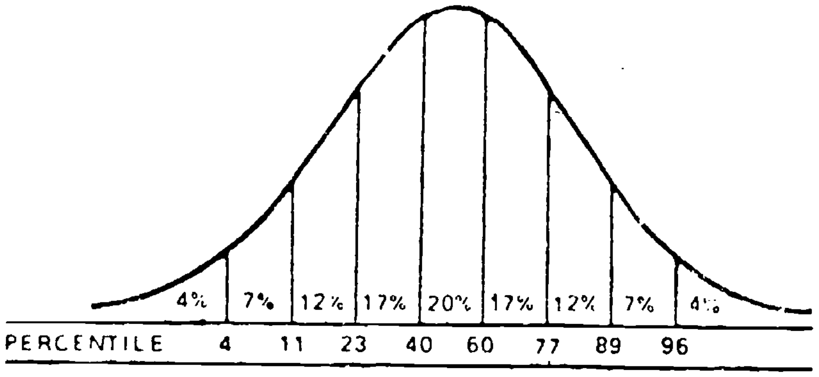
Criterion-referenced tests, such as the Survey of Essential Skills (SES), often report percent correct scores for specific skills which have been taught. This helps to show the degree of mastery a student has attained in that skill and whether additional instruction is needed.

Percentile Scores

Neither a raw score nor a percent correct score by itself is sufficient to tell whether the score is high or low. The comparison of an individual score with that of a group will greatly expand its meaning. This may be done within a classroom by looking at and ranking the scores of the whole class, or it can be done by referring to the norm group if it is a standardized, norm-referenced test.

One way of describing a student's score in relationship to the scores of other students is the relative rank of the particular score. The term percentile rank tells what percentage of the students in the norm group received the same raw score or lower raw score on the same test. For example, if a raw score of 42 correct answers on a specific test has a percentile rank of 70, it means that 70% of the students in the norm group got a raw score of 42 or less.

This percentile rank has meaning only in relation to the specific norm reference group. It does not apply to any other test or to any other group. The percentile rank is <u>not</u> the percent of correct answers. It is the percent of students in the norm group who earned the same or a lower raw score. Percentiles are used to report school data as well as individual student data. The percentile rank can be shown, relative to the normal curve, as follows:



The percentile ranks shown are simply the cumulative percentages of students scoring at or below a particular score - starting from the left (lower) end of the scale.

Scale Scores

Scale scores are expressed as numbers that may range from 0 to 999. A raw score is converted to a scale score by using a conversion table. Scale scores are especially appropriate for statistical purposes, because they can be added, subtracted and averaged across test levels which allows for direct comparisons among tests ana test levels. It is difficult to interpret an individual's performance using scale scores; therefore, they are primarily used to obtain derived scores such as percentiles and stanines.

The scale score is the basic score for CTBS, Forms U and V.

Grade Equivalent Scores

Test scores are sometimes stated as grade equivalent scores which indicate a grade level in years and months. The grade equivalent score is analogous to the average raw score for a particular grade level in the standardization sample. For example, a raw score of 38 with a grade equivalent score of 5.4 means that, in the norm group, 38 was the average score for students in the fourth month of the fifth grade.

If a student has a very high or very low grade equivalent score, relative to his actual grade placement, it does not mean that the student can do the work of a higher grade or should be demoted to a lower grade. It means simply that the student is performing very well for the grade level, or that the student needs to improve in the work at the grade level.

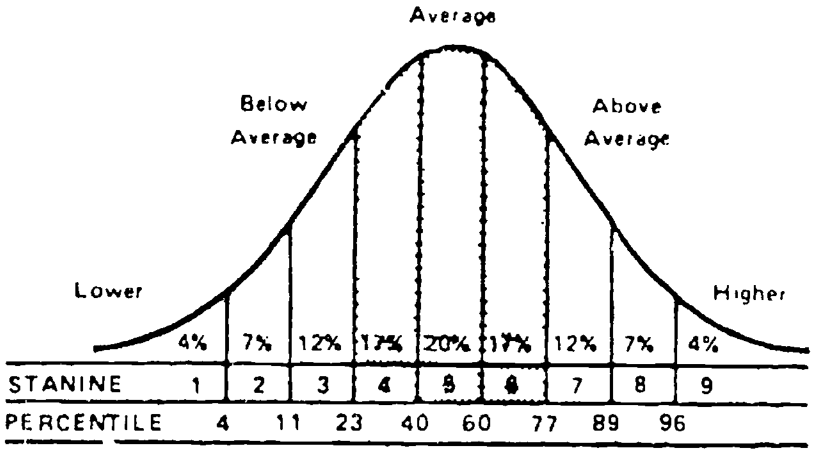
If we administer this test to a class which is in the fourth month of the fifth grade, it is not surprising that some members of the class do not score at exactly 38. We would expect some to be above the average and some to be below the average. However, if we find one score of 52, it does surprise us because our test manual indicates that this score has a grade equivalent of 6.4. It would be easy to conclude that the child who earned this score should be in the sixth grade. This is a common, but unfortunate, misinterpretation of the grade equivalent score.

The interpretation which is supported by the test score is that our high scoring child has performed on this test as well as the average pupil in the fourth month of the sixth grade would have performed on the same test. We must recognize that the score gives the impression of being more precise than it actually is and that an above average number of correct answers on a fifth grade test does not imply a knowledge of sixth grade curriculum materials which may not be measured on a fifth grade test.

Stanines

Stanines are another way to describe norm-referenced scores, as are percentiles, scale scores, and grade equivalent scores. The word stanine means a standard score scale of nine equal units. The scores range from 1 to 9 with a mean of 5. The limited range of values reduces the possibility of magnifying small differences between scores and are popular because they display real differences in a more readily understandable way than do most measures. But like the other norm-referenced scales, careful and cautious interpretation must be made of the scores.

Stanines are commonly used to group students for instruction - especially in large schools where there are many sections at each grade level, i.e., Stanines 1, 2, 3 = low; stanines 4, 5, 6 = average; and stanines 7, 8, 9 = high. This rough grouping is sufficient for instructional purposes.



Criterion-Referenced Scores

Since criterion-referenced tests are designed to show whether a student has mastered specific instructional objectives, the test score does not refer to a norm group or grade level, but, rather, shows the progress the student has made toward meeting the mastery or criterion level.

Scores may be reported as the percent of items answered correctly, as a range of mastery levels, as an absolute pass (mastery) or fail (non-mastery) score, or as a simple list of objectives with indications of which have been mastered. In all cases the mastery or criterion level of achievement will have been identified in advance so that score interpretation is clear.

Usually the individual's criterion test report shows the list of individual objectives or skills included in the test with possibly a mastery score for each objective and a total test mastery score.

Scores for Groups of Students

While the score of one student is easily described by a raw score, a percentile score, or even by a "passing" score, how are the scores of a group of 30 or of 30,000 students to be described? Two numbers, one representing the central tendency and the other representing the variability of the group, make it possible to describe the scores of the group as a whole.

Central Tendency

Scores for groups are usually reported as a single or "typical" score which represents the central tendency or average for the group. The two types of average scores most often used with standardized tests are mean scores and median scores. Thus, while individual student scores are reported as raw scores, percent correct, or percentiles, the average for a classroom, school, or district will be reported as mean raw score, mean percent correct, mean percentile, or as median raw score, median grade equivalent, or median percentile. Group scores may also be reported as Mode Scores.

Mean Scores

The group average which is called a mean score is computed by adding up all the student scores in the group and dividing the sum by the number of student scores. For example, a classroom might have the following raw scores on a test:

Students	Raw Scores
# 1	46
# 2	75
# 3	66
# 4	53
# 5	67
# 6	89
# 7	58
# 8	63
# 9	49
#10	62
#11	83
#12	62
#13	58
#14	67
#15	64
#16	62
#17	77
#18	91
#19	79
#20	73
#21	62
#22	68
#23	67
Total	1541

$$\begin{array}{r} 67 \\ 23\overline{)1541} \end{array}$$ (mean raw score)

By adding up all the student raw scores, one gets a sum of 1541. This figure is divided by 23 (the number of student scores) to arrive at the mean raw score. The mean is an especially useful average score because it is used as the basis for many statistical calculations.

Median Scores

Another score which is frequently used is the median. District CTBS scores are reported as median school or median district scores. The median is the middle score. In order to find the middle score, all scores for the group must be listed in numerical order. The score in the exact middle of the list is the median score. When one uses scores for the class shown on the previous page, the median score is 66 because 11 of the 23 scores fall above 66 and 11 fall below.

Raw Scores	
91	(highest score)
89	
83	
79	
77	
75	
73	
68	
67	
67	
67	
66	(middle or median raw score)
64	
63	
62	
62	
62	
62	
58	
58	
53	
49	
46	(lowest score)

Mode Scores

Another measure of central tendency is known as the mode. The mode is the score which occurs most often. The score, 62, occurs four times in the example above. The mode is, therefore, 62 because no other score occurs as frequently. A frequency distribution can have more than one mode.

Variability

The measures of central tendency tell us what is the "most typical" score for the group of students tested, but they do not tell us how the scores of individuals cluster or scatter around that average score. To describe the distribution of scores around the average, a measure of variability is used. The measures of variability most often encountered on test score reports are the range, the semi-interquartile range, and the standard deviation.

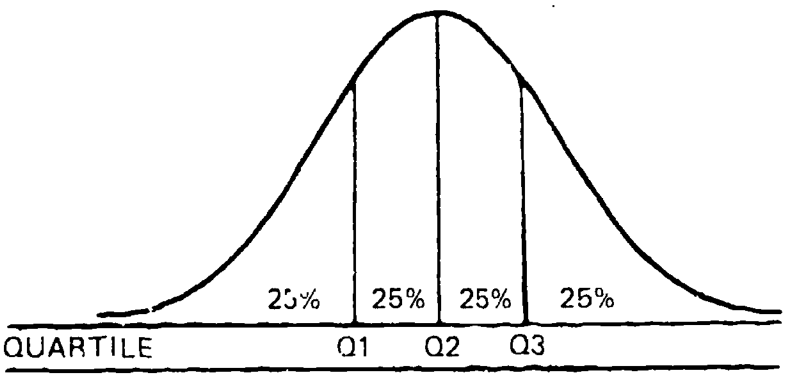
The Total Range

The total range is the indicator of variability that is the easiest to determine. It consists of the difference between the highest and lowest scores for the group. The range is a simple but unreliable estimate of the variability in a group of scores because only two scores are used to determine it. The remaining cores have nothing to do with the range measure.

Quartiles

Whenever the median is used as the measure of central tendency, the variability is shown by using the semi-interquartile range. Thi. statistic is defined as one-half the range between the 1st and 3rd quartiles.

The 1st quartile is computed by counting up from below to include the lowest or 1st quarter of the scores. It is given the symbol Q1. Counting down from above to include the highest or fourth quarter of the scores, one locates the 3rd quartile, or Q3. The median, which separates the second and third quarters of the distribution, is also called Q2. Note that the quartiles Q1, Q2, and Q3 are division points between quarters on the measuring scale.



In school or district group testing reports there often is information about the number or percent of students whose scores fall at or below a particular quartile point on the national norms. Thus a school in which 30% of the students score at or below Q2 can be compared with the national norm sample in which 50% of students score at or below Q2.

Standard Deviation

The most reliable and commonly used indicator of variability is the standard deviation. It is more difficult to compute but involves every score in the group and, therefore, is a more accurate estimate of the clustering or scattering of all scores in the group around the mean score.

In order to compute the standard deviation, one needs the mean score for the group. Each pupil score is subtracted from the group mean score;

these individual score deviations from the mean are then squared, and the squares summed. This sum is divided by the number of scores in the group (N), and the square root of the result is computed.

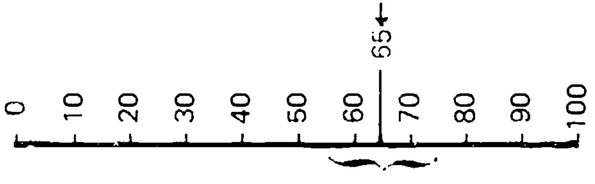
Interpretation of the standard deviation usually consists of stating the percentage of the individual pupil scores which would be included within the range from one standard deviation below the group mean to one standard deviation above the group mean. In a normal distribution, exactly 68.26% of scores, or about two-thirds, would be included in this range.

For example, if a test with a possible score of 100 points is administered to a group of 1500 students, how does one interpret a mean score of 65 and a standard deviation of 12? Assuming a statistically normal distribution of scores, one could expect that two-thirds of the students, 1000 of the 1500, scored between 53 and 77 on the test. Further, one-sixth of the students (250) will score above 77, and another one-sixth will score below 53.

If, in the above example, the standard deviation was 5 instead of 12, it would indicate that two-thirds of the students scored within a very narrow range, from 60 to 70 points.

With the thoughtful and appropriate use of a measure of central tendency and a measure of variability with group test scores, it is possible to understand the performance of the group as a whole and compare it with other groups who have taken the same test.

Mean score for group = 65



The range within which 2/3 of the scores are expected if S.D. = 12

8. HOW DEPENDABLE ARE THE SCORES?

Even when tests are administered under "standard" conditions and are properly scored, no scores can be relied upon as absolutely precise measurements. In asking how precise test scores are, one must first examine two ideas: *reliability* and *validity.*

A test is *reliable* to the extent that a student who receives a very high score on the test would receive a very high score on the same test administered again in the very near future. In other words, the reliability of a test relates to internal consistency of scoring. *Validity* means that the test provides an appropriate and relevant measure of the abilities it purports to measure and that, if the test is properly administered, a student's score will be representative of the student's knowledge.

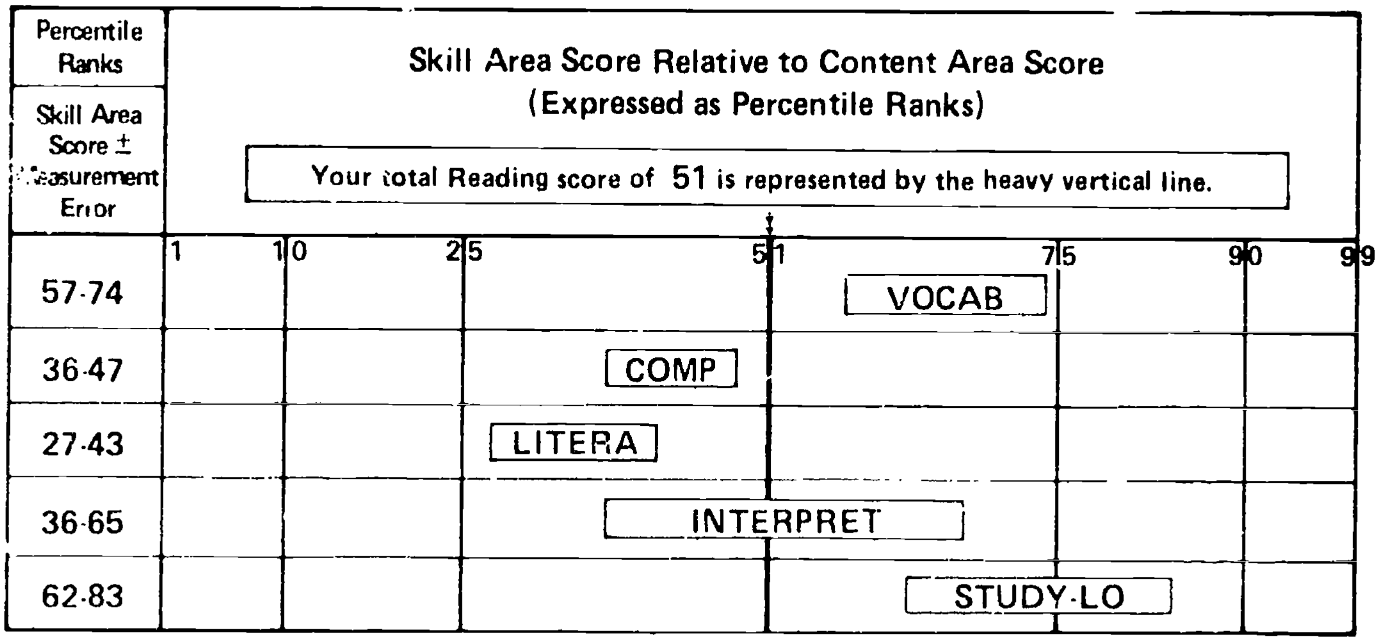
The question of dependability can now be discussed by considering that standardized tests really measure samples. First, it will be seen that a test only asks questions about limited areas of knowledge such as mathematics, history, or science. Within that limited area, only a sampling of what the student knows or does not know is assessed. Second, in establishing a norm or series of norms, only a small sample of all students are measured. For example, if norms are being established for fourth grade students, only a small portion of all fourth graders in the United States will be measured. Although there are methods to assure that these samples are statistically representative of the whole student population, it cannot be guaranteed that they are absolutely representative.

9. HOW ACCURATE ARE THE SCORES?

When we discussed the dependability of tests, we said that tests are based on samples. They measure samples of what the student knows and compare the result with the result obtained for a sample c" all the scudents in the population. There is always the possibility of a measuring error which would cause differences between what a test score implies about a student's knowledge and what the student actually knows. This error is called the <u>standard error</u> and can be determined by statistical methods.

Each standardized test has a published standard error which can be found and applied to a student's earned score. This means that if a student earns a score of 100 on a test with a standard error of 4, the student's true score should be expressed as lying within an interval 96 (100-4) to 104 (100+4). In other words, the earned score is really in the center of a group of scores described by the earned score and the standard error.

On the California Assessment Program (CAP) test reports, the standard error is clearly indicated for subtest skill areas. It is reported both as a number and, visually, as a band that extends on either side of the reported score.



10. NOW! WHAT DO TEST SCORES REALLY MEAN?

If a student scores well above average and if this performance confirms other evidence such as class recitation, homework quality, and other test results, it means that, in the areas tested, the student's progress is good. It may mean that the student could accept greater challenges or could benefit from an enriched or expanded program.

If a student scores substantially below average and if this confirms other evidence, including teacher observation, it probably indicates that the student needs help in the areas of weakness.

The same general considerations apply to grade equivalent scores and age equivalent scores. Small differences from the norm cannot be interpreted as being meaningful in any sense. Large differences from the norm provide a measure which can be used as one bit of evidence. When combined with other evidence, the total evaluation may be useful in identifying a student's strengths and weaknesses and in helping the teacher plan a suitable program to bolster strengths and overcome weaknesses.

School or school district scores can often be interpreted more definitively. For large groups, errors tend to cancel out, but the possibility of misinterpretation still exists. One additional very important fact must be considerea. If the size of the group being tested is quite small, the interpretation of the ranking of the school or district may be in error because a change in the score of one student may cause the school average to shift several points. In addition, schools with lower averages may have a number of excellent students while schools with higher averages may have many below average students. A school or district standing, like student rank, should be looked upon as one element of many which can be used in evaluating a school or district, but by no means the only element.

How are test scores used?

In the first section of this publication we stated that the purpose of testing was to improve the instructional program and to provide guidance for individual students.

Many students, parents, community groups, and educators question the value of testing programs. These critics cite inequities which include labeling (and mislabeling) students as slow learners, failing to account for possible test bias against some groups of students, and overemphasizing test scores as a basis for working with students and curriculum. These complaints appear to refute the reasons cited for testing and indicate the misuse of test data. Such complaints should not be directed at tests and test scores, per se, but at their improper use. Let us examine the complaints and try to place them in proper perspective.

Is there any value to a testing program?

Norms, whether national, regional, state, or local, represent the total population of students within which each of our students is measured. Our students will eventually be competing, in a highly complex and dramatically changing society, for career opportunities and for college admission. They will be in competition in a market closely parallel to the norm group. It is important to know how each of our students performs relative to the norm if we are to help them to obtain the best education possible and to use that education effectively.

Are some students mislabeled as "slow learners?"

Mistakes are probably made for a variety of reasons, but why is it necessary to label a student at all? Current practice is to avoid labeling students; however, student progress needs to be analyzed in order to determine how well each student appears to be progressing in each subject area. Without this analysis the teacher may fail to move a student fast enough to maintain interest, or a student may be moved faster than ability permits. In either case the chances for success are reduced. One group of students is lost through boredom; another group, through frustration.

Is there test bias which discriminates against some groups of students, such as racial or ethnic minorities?

The schools and test publishers have become more aware of these problems and have attempted to correct inequities by including persons representative of all backgrounds in the preparation and selection of tests. Clearly, tests that can be shown to be discriminatory against any group on any grounds should not be used to measure that group and probably should not be used at all. Efforts to design bias-free tests continue but have not yet been completely successful.

Is there overemphasis on test scores as a basis for student evaluation?

It is true that some educators and members of the public consider the test scores to be scientifically precise, sufficient, and conclusive data. These ideas will be of little help to the student or to the school. To use test data meaningfully, one must, at least, relate test scores to the following factors:

- past performance on similar tests, interest, attitude and personality profiles;
- motivation and aspirations;
- scholastic and attendance records;
- physical health;
- home and community cultural environment;
- mobility;
- teacher observation reports (anecdotal records).

Removed from this total context, the test score has limited value. Within the total framework, it can help crystallize what is known about a student to aid in placement, planning, guidance, and classroom instruction. No set of statistics can replace this total approach.

GLOSSARY

This glossary of terms used in educational measurement is intended for those who are not specialists in this field. The terms defined are the more common or basic ones which occur in measurement literature.

academic aptitude. See SCHOLASTIC APTITUDE.

achievement test. A test that measures the extent to which a person has accomplished tasks, acquired information, or mastered skills--usually as a result of planned instruction or training.

age norms. Values representing typical or average performance for persons of various age groups. Such norms are generally used in the interpretation of scholastic aptitude test scores.

aptitude. See SCHOLASTIC APTITUDE.

aptitude test. Tests which include those of general academic ability (commonly called mental ability or intelligence tests); those special abilities, such as verbal, numerical, mechanical, or musical; tests assessing readiness for learning; and prognostic tests which measure both ability and previous learning and are used to predict future performance.

arithmetic mean. See MEAN.

attitude test. A test which is intended to measure how an individual feels about specified things. Attitude about subjects such as school, school subjects, peers, or learning are often measured on a sliding scale so that the student may indicate a degree of positive or negative feeling toward the subject.

average. A general term applied to the various measures of central tendency. The three most widely used averages are the arithmetic mean (mean), the median, and the mode. When the term average is used without designation as to type, the most likely assumption is that it is the arithmetic mean.

bell-shaped curve. The graphic representation of a symmetrical frequency distribution. The normal curve is a special case of the bell-shaped curve family. See NORMAL CURVE.

bias (test bias). See CULTURE-FAIR TEST.

ceiling. The upper limit of ability that can be measured by a test. When an individual makes a score which is at or near the highest possible score, it is said that the test has too low a ceiling for him; he should be given a higher level of the test.

central tendency. A measure of central tendency provides a single most typical score as representative of a group of scores; the trerd of a group of measures as indicated by some type of average, usually the mean or the median.

criterion-referenced (content-referenced, objective-referenced) test. Terms often used to describe tests designed to provide information on the specific knowledge or skills possessed by a student. Such tests usually cover relatively small units of content and are closely related to instruction. Their scores have meaning in terms of what the student knows or can do, rather than in their relation to the scores made by some external reference group.

critique. A critical estimate or discussion evaluating or analyzing with knowledge or propriety, especially works of art or literature.

culture-fair test. So-called culture-fair tests attempt to provide an equal opportunity for success by persons of all cultures and life experiences. Their content must, therefore, be limited to that which is equally common to all cultures, or to material that is entirely unfamiliar and novel for all persons whatever their cultural background. See CULTURE-FREE TEST.

culture-free test. Ideally, a test that is free of the impact of all cultural experiences; therefore, a measure reflecting only hereditary abilities. Since culture permeates all of man's environmental contacts, the construction of such a test would seem to be an impossibility. Cultural bias is not eliminated by the use of non-language or so-called performance tests, although it may be reduced in some instances. In terms of most of the purposes for which tests are used, the validity (value) of a culture-free test is questioned; a test designed to be equally applicable to all cultures may be of little or no practical value in any.

deviation. The amount by which a score differs from some reference value, such as the mean, the norm, or the score on some other test.

diagnostic test. A test used to diagnose or analyze; that is, to locate an individual's specific areas of weakness and strength, to determine the nature of the weakness and strength, and, where possible, to identify causes. These tests are designed so that individual item responses reveal specific disabilities and deficiencies in achievement.

distribution. See FREQUENCY DISTRIBUTION.

error of measurement. See PROBABLE ERROR AND STANDARD ERROR.

essay examination. An examination requiring a written or oral response in the form of a discussion of the subject matter from a personal point of view. The essay is usually graded subjectively. See also SUBJECTIVE TEST.

examination. See TEST.

frequency distribution. A tabulation of the scores (or other attributes) of a group of individuals to show the number (frequency) of each score, or of those within the range of each interval.

grade equivalent (grade placement equivalent). The grade level for which a given score is the real or estimated average. Grade equivalent interpretation expresses obtained scores in terms of grade and month of grade, assuming a 10-month school year; e.g., 5.4 = fourth month of fifth grade.

grade norm. The average test score obtained by pupils of given grade placement. For example, if 38 was the average raw score for a norm group in the fourth month of the fifth grade, the grade norm is 38.

group test. A test that may be administered to a number of individuals at the same time by one examiner.

individual test. A test that can be administered to only one person at a time because of the nature of the test and/or the maturity level of the examinees.

informal test. See TEACHER-MADE TEST.

intelligence quotient (IQ). Originally, an index of brightness expressed as the ratio of a person's mental age to his chronological age, MA/CA, multiplied by 100 to eliminate the decimal.

intelligence test. See SCHOLASTIC APTITUDE TEST.

interest inventory. See INVENTORY.

internal consistency. Degree of relationship among the items of a test; consistency in content sampling.

interval (class interval). A group of numbers treated as a subset of a larger range of data from the highest number to the lowest number. The larger range is divided into class intervals for convenience of statistical treatment. For example, a test with scores ranging from 0 to 100 could be divided into intervals as follows: 0-9.9; 10-19.9; 20-29.9; ...90-99.9.

inventory. A questionnaire or checklist, usually in the form of a self-report, designed to elicit non-intellective information about an individual. Not tests in the usual sense, inventories are most often concerned with personality traits, interests, attitudes, problems, motivations, etc. See PERSONALITY TEST.

item. A single question or exercise in a test.

item analysis. The process of determining how well a given test item discriminates among individuals who differ in the characteristic being tested.

mean. The sum of a set of scores divided by the number of scores; the arithmetic mean.

measurement error. See PROBABLE ERROR and STANDARD ERROR.

measuring instrument. See TEST.

median. The middle score (mid point) in a distribution or set of ranked scores; the point (score) that divides the group into two equal parts; the 50th percentile. Half of the scores are below the median and half above it, except when the median itself is one of the obtained scores.

mental ability test. See SCHOLASTIC APTITUDE TEST.

mode. The score or value that occurs most frequently in a distribution.

multiple-choice item. A test item in which the examinee's task is to choose the correct or best answer from several given answers or options.

norm. Statistics that supply a frame of reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of pupils of various grades or ages in the test standardization group. Since they represent average or typical performance, they should not be regarded as standards or as universally desirable levels of attainment. The most common types of norms are percentile rank, grade equivalent, and stanine. Reference groups are usually those of specified age or grade.

norm group. The sample body of pupils upon which a test has been standaradized.

norm-referenced. Interpretation of a test score by relating it to the performance of the norm upon which the specific test was standardized.

normal curve. The bell-shaped curve, of infinite extent, which is the graphic representation of the normal distribution. See NORMAL DISTRIBUTION.

normal distribution. A frequency distribution which conforms to the theoretical ideal occurrence, by chance variation, of equally likely, mutually exclusive events. In such a distribution, scores or measures are distributed symmetrically about the mean, with as many cases up to various distances above the mean as down to equal distances below it. Cases are concentrated near the mean and decrease in frequency, according to a precise mathematical equation, the farther one departs from the mean. Mean and median are identical in a normal distribution. The assumption that mental and psychological characteristics are distributed normally has been very useful in test development work.

objective test. A test made up of items for which correct responses may be set up in advance; scores are unaffected by the opinion or judgment of the scorer. Objective keys provide for scoring by clerks or by machine. Such a test is contrasted with a subjective test, such as the usual essay examination, to which different persons may assign different scores, ratings, or grades.

objective-referenced test. See CRITERION-REFERENCED TEST.

percentile. One of the points which divide a distribution into 100 parts, each containing 1 percent of the cases. Percentile does not indicate the percent of correct answers. See PERCENTILE RANK.

percentile rank. The relative position of a test score with reference to the scores of the norm group. For example, if a raw score of 42 has a percentile rank of 70, it means that 70 percent of the norm group obtained a score of 42 or less. See PERCENTILE.

performance test. A test involving some motor or manual response on the examinee's part, generally a manipulation of concrete equipment or materials. Usually not a paper-and-pencil test.

(1) A performance test of mental ability is one in which the role of language is excluded or minimized, and ability is assessed by what the examinee does rather than by what he says (or writes). Mazes, form boards, picture completion, and other types of items may be used.

(2) Performance tests include measures of mechanical or manipulative ability where the task itself coincides with the objective of the measurement.

(3) The term performance is also used to denote a test that is actually a work-sample; in this sense it may include paper-and-pencil tests, as, for example, a test in bookkeeping, in shorthand, or in proofreading, where no materials other than paper and pencil may be required, and where the test response is identical with the behavior about which information is desired.

personality test. A test intended to measure one or more of the non-intellective aspects of an individual's mental or psychological make-up; an instrument designed to obtain information on the affective characteristics of an individual--emotional, motivational, etc.--as distinguished from his abilities. Personality tests include (1) the so-called personality and adjustment inventories which seek to measure a person's status by means of self-descriptive responses to a series of questions; and (2) rating scales which call for rating, by one's self or another, the extent to which a subject possesses certain traits.

population. The entire data from which a sample is drawn.

probable error. A statistic providing an estimate of the possible magnitude of error present in some obtained measure. The probable error of measurement is the amount by which an obtained score may differ from the hypothetical true score due to errors of measurement. The probable error of measurement is an amount that in about one-half of the cases the obtained score would not differ by more than one probable error of measurement from the true score. Probable error is seldom used in practice because the standard error is a more useful statistic. See also STANDARD ERROR.

profile. A graphic representation of the results on several tests, for either an individual or a group, when the results have been expressed in some uniform or comparable terms (standard scores, percentile ranks, grade equivalents, etc.). Usually developed for a group of sub-tests or sub-inventories within a principal test or an inventory. The profile method of presentation permits identification of areas of strength or weakness.

quartile. One of three points that divide the cases in a distribution into four equal groups. The lower quartile (Q1), or 25th percentile, sets off the lowest fourth of the group; the middle quartile (Q2) is tne same as the 50th percentile, or median, and divides the second fourth of cases from the third; and the third quartile (Q3), or 75th percentile, sets off the top fourth.

random sample. In research design, a sample of the members of some total population drawn in such a way that every member of the population has an equal chance of being included--that is, in a way that precludes the operation of bias on selection. The purpose in using a sample free of bias is, of course, the requirement that the cases used be representative of the total population if findings for the sample are to be generalized to that population. In a stratified random sample, the drawing of cases is controlled in such a way that those chosen are representative also of specified subgroups of the total population. See REPRESENTATIVE SAMPLE.

range. The difference between the highest and the lowest obtained score on a test for a specified group.

raw score. The first quantitative result obtained in scoring a test. Usually the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.

reliability. The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement. Reliability is usually expressed by some form of reliability coefficient or by the standard error of measurement derived from it.

representative sample. A sample that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purpose under investigation. In an achievement test norm sample, such significant aspects might be the proportion of cases of each sex, from various types of schools, different geographical areas, the several socioeconomic levels, etc. By using representative samples, one can develop very trustworthy test norms with relatively small norms groups.

sample. A limited number of cases drawn from a population in such a way that an analysis of the sample can be generalized to apply to the population.

scale score. A score expressed as a number from 0 to 999 on a single, equal-interval scale.

scholastic aptitude. The combination of native and acquired abilities that are needed for school learning; likelihood of success in mastering academic work, as estimated from measures of the necessary abilities. (Also called academic aptitude, intelligence, mental ability.)

scholastic aptitude test. A test which attempts to identify and measure scholastic aptitude.

standard deviation. A measure of the variability or dispersion of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, approximately two-thirds (68.3 percent) of the scores are within the range from one standard deviation below the mean to one standard deviation above the mean, and almost the entire normal distribution (99.7 percent) lies within three deviations from the mean.

standard error. A statistic providing an estimate of the possible magnitude of error present in some obtained measure.

(1) standard error of measurement: As applied to a single obtained score, the amount by which the score may differ from the hypothetical true score due to errors of measurement. The larger the standard error of measurement, the less reliable the score. The standard error of measurement is an amount that in about two-thirds of the cases the obtained score would not differ by more than one standard error of measurement from the true score. (Theoretically, then, it can be said that the chances are 2:1 that the actual score is within a band extending from true score minus 1 standard error of measurement to true score plus 1 standard error of measurement; but since the true score can never be known, actual practice must reverse the true-obtained relation for an interpretation.) See TRUE SCORE.

(2) standard error of the mean: When applied to group averages, standard deviations, correlation coefficients, etc., the standard error of the mean provides an estimate of the error which may be involved.

standard score. A general term referring to any of a variety of transformed scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. Standard scores are useful in expressing the raw scores of two forms of a test in comparable terms in instances where tryouts have shown that the two forms are not identical in difficulty; also successive levels of a test may be linked to form a continuous standard-score scale, making across-battery comparisons possible. Standard scores are usually based on the standard deviation of the test. For example, the College Board Exam score 600 is one standard deviation above the mean of 500. See NORM-REFERENCED.

standardized test (standard test). A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, under standard controlled conditions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Some further restrict the usage of the term standardized to those tests for which the items have been chosen on the basis of experimental evaluation, and for which data on reliability and validity are provided. These tests are usually commercially published for general use.

stanine. One of the steps in a nine-point scale of standard scores. The stanine (short for standard-nine) scale has values from 1 to 9, with a mean of 5 and a standard deviation of 2. Each stanine (except 1 and 9) is 1/2 standard deviation in width, with the middle (average) stanine of 5 extending from 1/4 standard deviation below to 1/4 standard deviation above the mean.

statistic. A single term or datum in a collection of statistics. See STATISTICS (2)

statistics. (1) A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data. (2) A collection of quantitative data.

subjective test. A test to which different scorers may assign different scores, ratings, or grades by exercise of judgment or opinion. A test which is subjectively evaluated. See ESSAY EXAMINATION and OBJECTIVE TEST.

teacher-made test. A test prepared by the teacher as contrasted with published tests which may be purchased or departmental tests prepared by committee action.

test A device for measuring specific characteristics or atttributes of individuals or groups by requiring the performance of a specified task or series of tasks.

true score. A score entirely free of error; hence, a hypothetical value that can never be obtained by testing, which always involves some measurement error. A true score may be thought of as the average score from an infinite number of measurements from the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the testings. The standard deviation of this infinite number of samplings is known as the standard error of measurement.

validity. The extent to which a test does the job for which it is used. This definition is more satisfactory than the traditional "extent to which a test measures what it is supposed to measure," since the validity of a test is always specific to the purposes for which the test is used. A test is valid to the extent that we know what it measures or predicts.

variability. The spread or dispersion of test scores, best indicated by their standard deviation.

References

Los Angeles Unified School District, Research and Evaluation Branch (1983). Handbook of Testing Programs in the Los Angeles Unified School District, 1983-84. (Research and Evaluation Branch Publication No. 432). (Available from Research and Evaluation Branch, 625-6362.)

Los Angeles Unified School District, Research and Evaluation Branch (1983). Norm-Referenced Test Results, CTBS Test Scores 1982-83. (Available from Research and Evaluation Branch, 625-6362.)

Los Angeles Unified School District, Research and Evaluation Branch (1983). Report on the District Testing Programs, 1982-83. (Research and Evaluation Branch Publication No. 433). (Available from Publications for Sale, 625-6628.)

Morris, Lynn Lyon, & Fitzgibbon, Carol Taylor (1978). How to Measure Achievement. Beverly Hills: Sage Publications.