DOCUMENT RESUME

ED 219 395 TM 820 398

AUTHOR Fortna, Richard O.
TITLE A Glossary of Measurement Terms Used in Title I Evaluation.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Office of Elementary and Secondary Education (ED), Washington, DC.
PUB DATE Jul 81
NOTE 13p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Testing; Evaluation Methods; *Glossaries; *Program Evaluation; Statistical Analysis; Test Interpretation; Test Reliability; Test Validity
IDENTIFIERS *Elementary Secondary Education Act Title I; Title I Evaluation and Reporting System

ABSTRACT

Measurement terms used in Title I evaluation are contained in this glossary. Several types of measurement techniques are identified and defined. Other measurement terms which are defined include those relating to validity, reliability, statistical analysis, test interpretation, and program effectiveness. (DWH)

DOCUMENT RESUME

ED 219 395 TM 820 398

AUTHOR Fortna, Richard O.
TITLE A Glossary of Measurement Terms Used in Title I Evaluation.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Office of Elementary and Secondary Education (ED), Washington, DC.
PUB DATE Jul 81
NOTE 13p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Testing; Evaluation Methods; *Glossaries; *Program Evaluation; Statistical Analysis; Test Interpretation; Test Reliability; Test Validity
IDENTIFIERS *Elementary Secondary Education Act Title I; Title I Evaluation and Reporting System

ABSTRACT

Measurement terms used in Title I evaluation are contained in this glossary. Several types of measurement techniques are identified and defined. Other measurement terms which are defined include those relating to validity, reliability, statistical analysis, test interpretation, and program effectiveness. (DWH)

A GLOSSARY OF MEASUREMENT TERMS USED IN TITLE I EVALUATION

JULY 1981

A GLOSSARY OF MEASUREMENT TERMS USED IN TITLE I EVALUATION

Richard O. Fortna
Assistant Director
Region II TAC

Achievement test--a test designed to measure a person's knowledge, skills, understandings, etc., in a given field taught in school. For example, a mathematics test or an English test.

Affective domain test--a test which is designed to measure a person's feelings or emotions.

Age norms--originally, values representing typical or average performance for persons of various age groups; most current usage refers to sets of complete score interpretive data for appropriate successive age groups. Such norms are generally used in the interpretation of mental ability test scores.

Alternate-form reliability--the closeness of correspondence or correlation between results on alternate (equivalent or parallel) forms of a test. The time interval between the two testings must be relatively short (no more than two to three weeks) so that the examinees themselves are unchanged in the ability being measured. (See Reliability coefficient.)

Aptitude--(1) a group of characteristics deemed to be symptomatic of an individual's ability to acquire proficiency in a given area; examples might be a particular art school subject, or vocational area; (2) ability measured by the amount of time required by the learner to acquire mastery of a task; thus, given enough time all students can conceivably attain such mastery.

Average--the sum of the measures, observations, magnitudes, items or scores divided by their number or frequency.

Basic skills test--a test intended to measure fundamental reading and computational skills which are the basis of later learning and achievement.

Ceiling--the upper limit of ability that can be measured by a test.

Ceiling effect--(1) the level above which a test ceases to distinguish between actual differences in the ability being tested; (2) a situation occurring when some members of a group cannot score as high as they are capable because the test is too easy. This results in an artificial restriction at the upper end of the score distribution. (See Floor effect.)

Central tendency (measure of)--any of various statistical measures which provides a single most typical value as representative of a group of measures, observations, magnitudes, items or scores. (See Average, Median, and Mode.)

Chance score--the score that one would expect if an examinee blindly guessed on every item.

Coefficient of correlation (r)--a measure of the degree of relationship between two sets of measures for the same group of individuals. The correlation coefficient most frequently used in test development and educational research is that known as the Pearson or product-moment. Correlation coefficients range from .00, denoting a complete absence of relationship, to +1.00 and to -1.00 indicating perfect positive or negative correspondence, respectively.

Coefficient of stability--a coefficient of reliability of the type based on correlation between test and retest, with an intervening period of time. (See Reliability coefficient.)

Comparison group model--also known as Model B, it is a variation of the control group design for evaluating Title I projects. No-treatment expectation is computed from the comparison group mean posttest NCE score.

Composite score--a score which combines several scores, usually by addition; often different weights are applied to the contributing scores to increase or decrease their importance in the resulting composite.

Correlation--(See Coefficient of correlation.)

Criterion--(1) a standard, norm or judgment selected as a basis for quantitative and qualitative comparison; (2) a standard by which a test may be judged or evaluated; a set of scores, ratings, etc., that a test is designed to measure, predict or correlate with; (3) a measure or standard that is used to make a judgment about the success of a project or prediction.

Criterion-referenced model--compares mastery levels achieved on specific objectives at the end of Early Childhood Education Title I projects with those at the beginning of the project.

Criterion-referenced test--tests designed to assess performance levels in relation to a set of well-defined objectives. 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. (See Norm-referenced test.)

Cutoff score--(See Pretest cutoff score.)

Decile--any one of the nine points (scores) that divide a distribution into ten parts, each containing one-tenth of all the scores or cases; every tenth percentile.

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 designed to provide in-depth information about specific weaknesses in a person's reading or mathematics skills that must be remedied before the person can be expected to make normal progress in their schoolwork.

Difficulty value--an index which indicates the percent of some specified group, such as students of a given age or grade, who answer a test item correctly.

Discrimination power--the ability of a test item to differentiate between persons possessing much or little of some trait.

Distractor--any incorrect choice (option) in a test item; also called a foil.

Distribution (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.

Early childhood education--educational experiences provided by the school at the pre-K, K, and 1st grade levels.

Empirically-normed test--any test which has norms based upon one or more of the commonly accepted methods for assuring a representative sample of students or individuals at the grade or age tested. Norms based upon "real" observations.

Equivalent form--two or more forms of a test which are so similar that they can be used interchangeably and yet are not identical; two or more test forms that yield about the same mean and variability of scores, and whose items are similar with respect to type, difficulty, distribution of item-test correlations and representative coverage of content.

Error of measurement--(See Standard error.)

Expanded standard score--a score system that links tests of different difficulty levels to the same scale; permits out of level testing and conversion to in-level norms.

Extrapolation--the process of inferring values of a variable in an unobserved interval from values within an already observed interval.

Floor--the lower limit of ability that can be measured by a test.

Floor effect--the level below which a test ceases to distinguish between actual differences in the ability being tested. (See Ceiling effect.)

Form--(See Equivalent form.)

Frequency distribution--(See Distribution.)

Functional level testing--testing with an appropriate difficulty level of a test for the group in question, avoids floor and ceiling effects; using a test level that distinguishes actual differences in the ability being tested. (See In-level test and Out-of-level test.)

Grade equivalent score--a converted score expressed in terms of a scale in which the grade is a unit of measurement and indicating the grade level of the group for which the score is typical or average; for example, a grade equivalent of 6.4 is interpreted as the fourth month of the sixth grade assuming a 10-month school year.

Grade norm--(1) the mean or median achievement of pupils in a given school grade on a given standardized test; (2) norms based upon the performance of pupils of given grade placement.

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

Halo effect--a positive bias in ratings arising from the tendency of a rater to be influenced in the rating of specific traits by the general impression of the person being rated.

Individual test--a test that can be administered to only one person at a time because of the nature of the test or the maturity level of the examinees.

In-level test--a test used with students in grades which correspond to the tests' nominal age/grade level. (See Out-of-level test and Functional level testing.)

Interpolation--in general, any process of estimating intermediate values between two known points; as applied to test norms, refers to the procedure used in assigning interpreted values, for example, grade or age equivalents, to scores between the successive average scores actually obtained in the standardization process.

Inventory test--an achievement test that attempts to cover rather thoroughly some relatively small unit of specific instruction or training.

Item--a single question or exercise in a test.

Item analysis--the process of evaluating single test items in respect to certain characteristics. It usually involves determining the difficulty value and the discriminating power of the item, and often its correlation with some external criterion.

JDRP--the Joint Dissemination Review Panel reviews evidence of effectiveness claimed for any educational program and supports dissemination of those found to be exemplary.

Level--a designation for a test or tests in a battery or series suited to a particular rank or plane of ability or achievement. In most achievement tests, typically grade in school.

Locator test--short tests accompanying some test batteries or series which can be used to help determine an appropriate test level for the individual or group to be tested. (See Level.)

Mastery level--a standard of performance on criterion-referenced or mastery tests.

Mastery test--a test designed to show whether a student has successfully performed all the tasks specified in a given learning objective.

Mean--(See Average.)

Median--the middle score in a distribution or set of ranked scores; the point (score) that divides the group into two equal parts; the 50th percentile; a measure of central tendency. (See Average and Mode.)

Minimum competency test--a test designed to allow an individual to demonstrate the acquisition of a skill at the lowest possible level that is deemed acceptable.

Mode--the score or value that occurs most frequently in a distribution; a measure of central tendency. (See Average and Median.)

Models A, B, C--(See Norm-referenced model, Comparison group model, and Regression model.)

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.

N--the symbol commonly used to represent the number of cases in a group.

Normal curve equivalent (NCE)--a normalized, standard-score scale for reporting achievement data referenced to the national population at each grade level. The scale has a mean of 50 and a range from 1 to 99.

Normal distribution--a distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance; scores or measures are distributed symmetrically about the mean, with as many cases at various distances above the mean as at equal distances below it and with cases concentrated near the mean and decreasing in frequency the further one departs from that average, according to a precise mathematical equation.

Norm-referenced model--usually referred to as Model A, one of the Title I evaluation models. In this model the treatment effect is defined as the difference between the treatment group NCE post-test score and the treatment group NCE pretest score. Publishers' national norms are used to estimate no-treatment expectation, hence the name norm-referenced model.

Norm-referenced test--an examination for which an individual's score indicates the relationship of the individual's performance to that of a specified norm group.

Norms--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 standardization group for the test.

Since they represent average or typical performance, they should not be regarded as standards or as universally desirable levels of attainment.

No-treatment expectation--an estimate of what the Title I group posttest performance would have been without special instructional intervention. (See Project impact.)

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. Contrasted with a "subjective" test, such as an essay examination, to which different persons may assign different scores, ratings or grades.

Out of level test--a test used for students in grades either above or below the test's nominal age/grade level. When such tests are used they must have scores which allow a student's performance to be related to his or her own age/grade level. (See Expanded standard scores, Functional level testing and In-level test.)

Parent advisory council--mandated by the law and the regulations, these advisory groups must be involved in the planning, implementation, and evaluation of the Title I project.

Percentile--one of the 99 point scores that divide a ranked distribution into groups or parts, each of which contains 1/100 of the scores or persons; the points are located to coincide with the obtained scores below which in each division 1/100 of the cases fall; thus, a score coinciding with the 30th percentile is regarded as equaling or surpassing that of 30 percent of the persons in the group. "Percentile" has nothing to do with the percent of correct answers an examinee makes on a test.

Percentile rank--the expression of an obtained test score in terms of its position within a group of 100 scores; the percentile rank of a score is the percent of scores equal to or lower than the given score in its own or in some external reference group.

Performance test--(1) broadly, any test intended to measure actual accomplishment rather than potential ability or aptitude, regardless of how the subject is instructed to respond; (2) a test involving some motor or manual response on the examinee's part; usually not a paper-and-pencil test.

Posttest--a test given at the conclusion of an educational project or treatment to determine post-treatment status of the examinee or group in regard to some skill, aptitude or achievement. (See Pretest.)

Power test--a test intended to measure level of performance unaffected by speed of response; hence, one in which there is either no time limit or a very generous one. Items are usually arranged in order of increasing difficulty.

Practice effect--(1) a change that follows practice in taking a test; usually an increase in the score of an individual when the same test is taken more than once; (2) the apparent gain in accomplishment resulting from using the same test on two or more occasions.

Pretest--(1) a test given in order to determine the status of the examinee or group in regard to some skill, aptitude, or achievement, as a basis for judging the effectiveness of subsequent treatment; (2) a tryout of some test in advance of its regular use. (See Posttest.)

Pretest cutoff score--a score which divides an original intact group of students into treatment and comparison subgroups (in the Regression model) in such a way that all students having pretest scores lower than the cutoff value receive the treatment and all students scoring above the cutoff value are excluded from the treatment.

Project impact--difference between actual performance and estimated no-treatment posttest performance of Title I group receiving treatment.

Project vectors--characteristics and impact of Title I projects at grades 2,6, and 10.

Psychological test--(1) an examination which determines the relative strength of intellectual ability in terms of standard; (2) an examination which measurers some nonphysical aspect of human behavior.

Psychomotor test--an examination to measure the motor effects (e.g., dexterity, physical movement, etc.) of a person's mental or cerebral processes.

Quartile--one of three points that divide the cases in a distribution into four equal groups. The lower quartile (Q1), or 25th percentile marks the lowest quarter of the group; the middle quartile (Q2) is the same as the 50th percentile or median; and the third quartile (Q3), or 75th percentile, sets off the top quarter.

Random sample--a sample of the members of some total population selected in such a way that every member of the population has an equal chance of being included in the sample.

Range--for some specified group, the difference between the highest and lowest obtained score on a test, thus a very rough measure of spread or variability.

Raw score--a quantitative result obtained in scoring a test; usually the total number of right answers.

Readiness test--a test that measures the extent to which an individual has achieved a degree of maturity or acquired certain skills or information needed for successfully undertaking some new learning activity. Readiness tests are also classified as prognostic tests.

Regression effect--tendency of a predicted score to be nearer to the mean of its distribution than the score from which it is predicted is to its mean. Because of the effects of regression, students making extremely high or extremely low scores on a test tend to make less extreme scores, i.e., closer to the mean, on a second administration of the same test or on some predicted measure.

Regression (line)--if two paired lists of numbers, say pretest scores and posttest scores, are plotted in two dimensions, say pretest horizontally and posttest vertically, then there is exactly one straight line that can be drawn through the plot so that it passes closest to the means of all those sets of posttest scores that correspond to each pretest score. On the average, for all pretest scores, this is the best straight-line fit to the observed posttest scores.

Regression model--also known as Model C, or the special regression model. In this model the treatment effect is defined as the difference between the treatment group posttest NCE score and the posttest NCE score estimated from the comparison group regression line.

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 coefficient--the coefficient of correlation between scores on two forms of a test, between scores on two administrations of the same test, or between scores on two halves of a test, properly corrected. (See Alternate form reliability and Coefficient of stability.)

Reliability of classification--parallel forms reliability for criterion-referenced tests determined by summing the proportion of people classified as masters and nonmasters by both tests.

Representative sample--a sample that corresponds to or approaches the population of which it is a sample with respect to characteristics important for the purposes under investigation. (See Technical standards.)

Sample--a finite part of a statistical population whose properties are studied to gain information about the whole.

Scale--(1) a series of numbers, such as norms, percentile scores, grade equivalents, etc., the values of which take significance from their derivation; (2) a test having items arranged in order of difficulty.

Significance, statistical--the property of having low probability of occurrence on the basis of chance alone, thereby likely occasioned by factors other than chance; not necessarily synonymous with practical significance.

Skewed distribution--a distribution that departs from symmetry or balance around the mean, i.e., from normality. Scores pile up at one end and trail off at the other.

Slope--the inclination or steepness of a line when it is viewed from left to right, a line's slope is the amount of vertical rise per unit of horizontal run.

Speeded test--a form of a test in which performance is measured by the number of tasks performed in a given time; it is assumed that the items are uniform in difficulty; primarily intended to yield a rate score that is not affected by other dimensions of pupil performance.

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 (S.D.). For a normal distribution, about two thirds (68.3 percent) of the scores are within the range from one S.D. below the mean to one S.D. above the mean.

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. The simplest type of standard score, known as a Z-score, is an expression of the deviation of a score from the mean score of a group in relation to the standard deviation of the scores of the group. Thus,

$$\text{standard score}(Z) = \frac{\text{raw score}(X) - \text{mean}(M)}{\text{standard deviation (S.D.)}}$$

Standardized test--a test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Other restrictions include those whose items have been chosen on the basis of experimental evaluation, and for which data on reliability and validity are provided. Others add "commercially published."

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. Each stanine (except 1 and 9) is 1/2 S.D. in width.

Survey test--a test that measures general achievement in a given area, usually with the connotation that the test is intended to assess group status, rather than to yield precise measures of individual performance.

Sustained gain--measurement to determine whether the student achievement fostered through Title I service is continued beyoned the end of the Title I project, that is, at least 12 months after the pretest.

Technical standards--mandated standards for evaluation of Title I project effectiveness. LEA must explain in its application how its evaluation will be consistent with these standards. They are: (1) representativeness of evaluation findings; (2) reliability and validity of evaluation instruments and procedures; (3) procedures which will minimize error; and (4) valid assessment of achievement gains in reading, mathematics and language arts.

TIERS--acronym for the Title I Evaluation and Reporting System.

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 individual during the testings.

Type I error--rejecting as false a statement or condition which is true.

Type II error--not accepting a true difference in condition as true.

Validity--the extent to which a test does the job for which it is used. Generally classified into four types:

Face validity--extent to which a test appears to measure what it is intended to measure.

Content validity--the extent to which the content of an achievement test represents a balanced and adequate sampling of the universe of content in which the test is intended to measure achievement. It is best evidenced by a comparison of the test content with courses of study, instructional materials, and statements of instructional goals.

Predictive validity--the extent to which a test accurately indicates future learning success in the area for which it is used as a predictor. Evidence of predictive validity is shown by correlations between scores on the test and future criterion measures of success. High school aptitude test scores highly correlated with first year college grade point averages is one example of predictive validity.

Concurrent validity--the extent to which scores on a test are in agreement with some given criterion measure. There is usually no significant time interval lapse between the time the test scores and criterion measures are obtained. For example, teacher ratings of reading ability made at about the same time as the posttest is administered could be correlated to obtain an index of concurrent validity.

Validity coefficient--the coefficient of correlation between a criterion variable and one or more independent variables that purport to measure or are used to predict the criterion.

Value added model--uses the relationship between test score and age and the information it contains about natural growth to project a "no-treatment" expectation for Early Childhood Education Title I project participants. (See Criterion-referenced model.)

Variance--a measure of variability equal to the square of the standard deviation; the average of the squared deviations from the mean.

Weighted mean--an average of a list of averages or means that takes into account the number of cases that entered the computation of each average.

Bibliography

1. Combined Glossary: Terms and Definitions from the Handbook of State Educational Records and Reports Series, NCES, USDHEW, Washington, DC, 1974.

2. Test Service Notebook 13 - A Glossary of Measurement Terms, Harcourt Brace Jovanovich, Inc., New York, 1976.

3. Good, C. V., Ed. Dictionary of Education, 3rd Ed., McGraw-Hill Book Co., New York, 1973.