Glossary of Evaluation and Statistical Terms

by
Richard J. McCowan
Glossary of Evaluation and Statistical Terms
The Center for Development of Human Services is a continuing education enterprise of the Research Foundation of the State University of New York and a unit of the Graduate Studies and Research Division at Buffalo State College (SUNY). Funding is provided by the New York State Office of Children and Family Services. © 1999 Research Foundation of SUNY/Center for Development of Human Services. All rights reserved.

Center for Development of Human Services
Robert N. Spaner - Chief Administrative Officer
bobs@bsc-cdhs.org
Richard J. McCowan, Ph.D. - Director, Research & Evaluation Group
dickm@bsc-cdhs.org
Buffalo State College (SUNY)
1695 Elmwood Avenue
Buffalo, New York 14207-2407
Tel.: 716.876.7600
Fax: 716.796.2201
http://www.bsc-cdhs.org
Evaluation & Statistical Terms

Foreword

This glossary contains a list of terms that are important in the fields of evaluation and statistics. It provides a handy reference for people who are unfamiliar with evaluation and statistical techniques and terminology.

The definitions are intentionally brief and simple. Consequently, the glossary does not describe how to perform the statistical analyses or conduct the evaluations. This information is available from other sources such as those listed in the reference section.

Scriven (1991) compiled an Evaluation Thesaurus that provides an expanded discussion for most of these terms. He also describes additional terms and concepts that were not included. The International Encyclopedia of Education Evaluation edited by Walberg and Haertel (1990) contains 150 short, but comprehensive, discussions of evaluation methods.

Many excellent statistical texts have been published. Kerlinger’s (1973) Foundations of Behavioral Research is a classic reference. Statistical Methods in Education and Psychology by Glass and Hopkins is comprehensive and clearly written. Miller’s (1991) Handbook of Research and Design and Social Measurement contains information not readily found in other sources. The best Internet site for evaluation is the ERIC Clearinghouse on Assessment and Evaluation (ERIC/TM). The following information can be used to contact this site.

University of Maryland—College Park
1129 Schriver Hall
College Park, MD 20742
Phone: 301/405-7449
Toll Free: 800/G04-ERIC
E-mail: ericae@ericae.net
Gopher: gopher.cua.edu, Special Resources
URL: http://ericae.net
Evaluation Terms

A

Achievement test: objective examination that measures skills or knowledge in subjects such as reading, spelling, or mathematics.

Accountability: documenting the acceptability of agreed-on criteria.

Adaptive testing: computer-assisting testing in which item presentation is dictated by the examinee’s responses to previous items.

Affective: related to attitudes, interests, values, feelings, preferences, likes/dislikes, and satisfaction.

Age Norms: scores representing average performance of people of age groups.

Alternative: choice or option in a multiple-choice item.

Alternate form reliability: (see “Reliability”).

Anchoring: procedure used to establish extreme response positions on attitude scales.

Answer key: list or correct or expected answers.

Aptitude: combination of abilities that indicate potential ability to learn.

Aptitude test: predictor of cognitive achievement prior to selection for a course or program.

Assessment: term often used as a synonym for “evaluation.” It is the systematic evaluation of program outcomes or the effects of an intervention on people. Some evaluators use “assessment” in reference to procedures that use testing and quantification rather than judgment.

Attribute: characteristic of persons or things.

Attitude: (see “Specific objective”).
**Attitude test:** measures the tendency to respond favorably or unfavorably towards groups, institutions, or objects.

**Attitude scale:** self-report instruments designed to assess affect, attitudes, and feelings.

*Thurstone scales*: equal-appearing interval scales developed by collecting a large number of items that express varying degrees of negative, neutral, and positive feelings about a variable. Competent judges evaluate items, and each item is given a scale value using mean ratings of judges to indicate the strength of the response.

*Likert scales*: scales typically rated on a 5-point or 7-point continuum with categories ranging from very low to very high (or strongly disagree to strongly agree).

*Semantic differential*: bipolar adjectives representing evaluation (good-bad), potency (strong-weak), and activity (fast-slow) are rated using a 7-point scale.

*Guttman scale*: cumulative scale in which a small set of homogeneous, unidimensional items measures a single attitude. Items are arranged so that a response indicates approval on all preceding items.

**Average:** measure of central tendency.

*Mean*: sum of a set of scores divided by the number of scores.

*Median*: midpoint of a set of scores.

*Mode*: score that appears most frequently in a set of scores.

**Battery:** group of tests administered to a population.

**Behavioral objective:** (see “Specific objective”).

**Benefits:** valued outcomes or processes. (see “Cost-benefit”).
**Bias**: quality of a test that misrepresents what is measured. In evaluation, “bias” means prejudice, and its antonyms are fair, objective, or impartial.

**Ceiling**: upper limit of ability measured by a test.

**Chance score**: score attained when individuals randomly respond to all items on a test.

**Chunking**: cognitive operation in which people cluster from 5 to 9 items or concepts into single units to facilitate memory.

**Cluster sample**: selected units that occur naturally in the population.

**Cognitive**: mental abilities such as recall, comprehension, problem solving, and synthesis.

**Competence**: related set of attitudes, skills, and knowledge.

**Composite measure**: combining scores from different tests to predict behavior.

**Concurrent validity**: (see “Validity”).

**Confidence interval**: range of values within which a population parameter is estimated to lie.

**Confidence level**: estimated probability that a parameter lies within a given confidence interval.

**Construct validity**: (see “Validity”).

**Content validity**: (see “Validity”).

**Convergent validity**: (see “Validity”).

**Control group**: group of randomly assigned subjects to who no treatment is assigned.

**Correction for guessing**: reduction in score for wrong answers sometimes applied in scoring multiple-choice questions.
**Correlation**: measure of the extent to which two variables are related.

**Cost-benefit**: analysis that answers the question: Is this program worth its cost, or which option has the highest benefit/cost ratio?

**Criterion**: measure used as a standard of performance.

**Criterion-referenced test**: designed to measure content specified by behavioral objectives.

**Criterion-related validity**: (see “Validity”).

**Cross-validation**: determining if a decision based on one data set is correct by applying the decision process to an independent, but relevant set of data.

**Culture-fair test**: test in which the influence of biasing factors such as sex or race is eliminated.

**D**

**Dependent variable**: variable that is affected by the independent variable (e.g., a test or criterion measure).

**Deviation**: amount by which a score differs from a reference value.

**Diagnostic test**: intensive evaluation with detailed coverage of a specific area. The test assesses specific individual learning needs.

**Difference score**: difference between pretest and posttest scores.

**Differential weighting**: assigning different scoring weights to test items.

**Difficulty level**: proportion of individuals who respond correctly to an item.

**Dispersion**: distribution of values around a mean or median.

**Discriminant validity**: (see “Validity”).
**Discrimination**: extent to which a test item distinguishes high achieving from low achieving trainees.

**Distracter**: alternative on a multiple-choice item.

**Domain**: set of related competencies, or population specifications for expected outcomes, behaviors, or learning.

**Domain-referenced test**: measures performance against a well-defined set of tasks or body of knowledge.

**E**

**Equivalent forms**: two or more test forms with similar content, difficulty, mean scores, and variability.

**Evaluation**: process by which quantitative and qualitative data are analyzed to determine the program value, worth, or effectiveness.

**Expressive objectives**: important consequences that cannot be expressed in behavioral terms.

**External criterion**: external standard of excellence against which trainee test performance is compared.

**F**

**Face validity**: (see “Validity”).

**Formative evaluation**: conducted periodically to provide on-going feedback to help staff improve a program. (see “Summative evaluation”).

**Frequency distribution**: tabular, sequential arrangement of scores showing the numbers of person who obtained different scores.

**G**

**Generalizability**: extent to which a test produces comparable results across different settings and populations.
**Goal-free evaluation:** investigation of all consequences of a program, not only those specified by writing objectives.

**Grade equivalent:** estimated grade level that corresponds to a given score.

**Guttman scale:** (see “Attitude scale”).

**H**

**Halo effect:** biased, inflated ratings on specific items based on an overall general impression of an individual.

**Hawthorne effect:** tendency of subjects to change behavior simply because they are part of a new experience, rather than because of the experimental treatment.

**High-stakes test:** test results that have a significant effect on individuals or groups.

**Histogram:** graph with vertical bars of different lengths in which the height of each bar corresponds to the frequency of scores.

**Hypothesis:** conjectural statement about the relationship that exists among predefined variables.

**I**

**Independent variable:** variable that is manipulated to determine the effect on a dependent variable (i.e., treatment or experimental variable).

**Informal test:** nonstandardized, teacher-made test that gives an approximate measure of individual performance.

**Intact sample:** existing group of nonrandomly assigned subjects.

**Interpolation:** process of estimating intermediate values between two known points.

**Interview:** planned procedure used to obtain oral information in a face-to-face situation.
**Inventory:** list of criteria used to assess the absence or presence of specific attributes relevant for a given purpose.

**Item:** individual test question.

**Item analysis:** procedure used to determine item characteristics, such as difficulty and discrimination.

**Item response theory:** (see “Latent trait theory”).

**Item sampling:** procedure used in test standardization in which subgroups take subsets of items.

**J**

**John Henry effect:** tendency of a control group to behave differently (e.g., work harder) because they realize they are the control group.

**Judgmental sample:** nonprobability sample selected on the basis of judgment rather than randomly.

**K**

**Knowledge:** (see “Specific objective”).

**L**

**Latent trait theory:** statistical procedures used to relate individual test scores to an estimated hypothetical latent trait.

**Level of significance:** probability that an observed relationship is a function of chance or sampling error. If a test is significant at the.05 level, it means that differences of that magnitude would occur by chance only 5 times out of 100.

**Likert scale:** (see “Attitude scale”).

**Longitudinal study:** collection of data on a sample at different points in time.
Mastery-level: acceptable performance level based on specific criteria related to individual, rather than group, achievement.

Mastery test: test the extent to which a person has mastered a specified set of objectives or met minimum requirements (usually a criterion-referenced test).

Mean: (see “Average”).

Measurement: assignment of numbers to attributes of people, events, or objects.

Measurement-driven instruction: use of criterion-referenced measures to dictate instructional experiences, particularly in relation to high-stakes testing.

Median: (see “Average”).

Minimum competency test: measures of basic skill required for promotion or job retention.

Mode: (see “Average”).

Multiple-choice item: test question consisting of a “stem” and several distracters, one of which is the correct answer.

Multiple correlation: relationship between one variable and the weighted sum or two or more other variables.

Multiple regression: combining two or more predictors to estimate a single criterion measure.

Multiple-response item: type of multiple-choice item in which two or more choices may be correct.

Multi-stage sample: clusters are selected and individuals are selected from each cluster.

Multivariate analysis: analysis of the simultaneous relationships among several variables (e.g., age, gender, and education on income).
Norm: performance standards established by a reference group that describe average or typical performance.

Normal curve: mathematically defined, bell-shaped curve in which 68% of the scores fall between z scores of +1.00 and -1.00.

Normal curve equivalent: standard scores with a mean of 50 and a standard deviation of 21.06.

Norm-referenced test: objective test standardized on a group of individuals whose performance is evaluated in relation to the performance of others.

Norms: statistics that describe the test performance of specified subgroups, which represent a larger population.

Objective test: test with clear, unambiguous scoring criteria.

Ogive: smooth curve resulting from the plot of cumulative frequencies.

Omnibus test: test in which items measuring varied mental operations are combined into a single test with a single score.

Option: (see “Distracter”).

Operational definition: concrete definition of a variable in terms of the operations that will be used to measure it.

Outcome: post-treatment effects based on selected criteria (e.g., time: immediate or long-term; impact: agency, community, client).

Parameter: measure taken from a population.
**Pareto Principle**: 80/20 rule or the principle of the vital few and trivial many that asserts that 80% of significant achievement is accomplished by 20% of those involved.

**Percentile**: one of 99 point scores that divide a ranked distribution into groups each of which is 1/100 of the scores, or a point below which a certain percentage of scores fall.

**Performance test**: test composed of motor or perceptual items designed to evaluate general intelligence or aptitudes.

**Pluralistic assessment**: approach to assessment that attempts to respond to cultural diversity.

**Population**: all individuals (or things) whose characteristics will be measured.

**Portfolio assessment**: collection and evaluation of student created performances according to expected criteria.

**Power test**: untimed test intended to measure level of performance rather than speed.

**Practice effect**: effect of previous experience with a test on a later administration of the same or a similar test.

**Predictive validity**: (see “Validity”).

**Published test**: commercially produced, copyrighted test available for public use.

---

**Q**

**Q-sort**: technique in which subjects sort statements into piles according to specific criteria (e.g., trainers sort multiple choice questions into three piles representing acceptable, unacceptable, and questionable).

**Quartile**: one of three points that divide a distribution into four equal groups (i.e., 25th percentile; 50th percentile; 75 percentile).
Random sample: sample in which each unit has the same probability of being selected.

Range: measure of dispersion consisting of the highest and lowest score in a distribution.

Rating scale: numerical or descriptive scale that allows subjective assessments on criteria presented as a scale.

Raw score: number of items that are answered correctly.

Readiness test: test that measures the extent to which a person is prepared (by maturity or skill) to undertake a new activity.

Recall item: items that require an examinee to recognize a correct answer from several given distracters.

Regression effect: tendency for a score to be closer to the mean than the score from which it was predicted or derived.

Relevance: extent to which items actually measure specific objectives.

Representative sample: sample selected in a random, unbiased manner.

Reliability: extent to which a test is consistent (dependable, error-free, and stable) when administered to the same individuals on different occasions.

Stability: administer the same test to the same people at different times.

Equivalence: administer different test forms to same group at the same time.

Stability & Equivalence: administer different test forms to the same people at the same time.

Split-half technique: compare odd to even numbers or a randomly selected 50% to the remaining 50%.
**Cronbach’s Alpha:** compare the variance of each item to total test variance.

**Decision-consistency:** average of squared deviations from established mastery level (for criterion-referenced tests).

**Respondent:** person who responds to a survey.

**Response rate:** number of persons participating in a survey divided by the total number included in the sample.

**Representative sample:** sample that accurately represents the population from which it was selected.

**S**

**Sample:** limited number of individuals selected from a population.

**Sampling frame:** list of individuals (or things) from which a sample is selected.

**Screening:** fast measurement for large populations to identify individuals who deviate in specific areas.

**Semantic differential:** (see “Attitude scale”).

**Skill:** (see “Specific objective”).

**Social desirability:** source of invalidity on self-report items whereby respondents answer in a social desirable way.

**Sociogram:** diagram of interpersonal relationships within a group showing friendship choices and rejections.

**Snowball sample:** nonprobability sample, which begins with several subjects who recommend additional subjects.

**Specific determiners:** ambiguous words that give clues to answering multiple-choice items.
**Specific objective:** describes in measurable terms what trainees must do to demonstrate performance.

**Attitude:** state of mind or a feeling.

**Skill:** proficiency requiring use of voice, senses, or body.

**Knowledge:** specific information that is perceived or learned.

**Specific determiner:** clue in an item that suggests a distracter is correct or incorrect.

**Specimen set:** sample set of testing materials from a commercial test publisher including the test, manual, and administration procedures.

**Spiral omnibus test:** test with different item formats in which item difficulty increases as the respondent progresses through the test.

**Standard deviation:** measure of variability or dispersion of scores around the mean.

**Standardized test:** test administered to a population after which standardized scores, percentiles, and equivalent scores are produced to compare individual scores to the norm group.

**Standard score:** score expressed as a deviation from a population mean.

**Stanine:** standard score obtained by dividing the normal curve into nine segments ranging from 1 to 9 with groups spaced in half-sigma units.

**Statistic:** measure taken from a sample.

**Stem:** question or open-ended statement in a multiple-choice item.

**Stratified sample:** population is divided into sub-populations and the sample is drawn randomly from each stratum.

**Subtest:** collection of items in a test that have distinct similar characteristics and yield a separate score.
**Summative evaluation**: conducted after completion of a program for the benefit of an external audience or decision-maker.

**Systematic sample**: sample selected by taking every \( n \)th unit.

**T**

**T-score**: standard score with a mean of 50 and a standard deviation of 10.

**Taxonomy**: classification plan of hierarchically arranged elements.

**Test**: systematic procedure for gathering data to make group and individual comparisons.

**Test-wiseness**: competence in test-taking that allows examinees to perform well on tests, particularly those created by inept test constructors.

360 degree evaluation: evaluation approach that collects information from all significant areas related to the evaluation (e.g., a broad circle of sources or “360 degrees”).

**Thurstone scale**: (see “Attitude scale”).

**True score**: average score on an infinite series of administrations of the same or equivalent tests.

**U**

**Univariate analysis**: analysis of the effect of a single independent variable on a dependent variable.
Validity: extent to which a test measures what it was intended to measure.

Concurrent: extent to which measures correlate with a criterion when both are obtained at approximately the same time.

Construct: extent to which test scores measure theoretical constructs.

Content: item content corresponds to specific objectives and curriculum.

Convergent: extent to which two or more sets of measures are highly correlated and measure the same trait.

Criterion-related: correlation between predictor and criterion measure.

Discriminant: extent to which a trait does not correlate with irrelevant variables.

Face: extent to which a test appears relevant, important, and interesting.

Predictive: extent to which a test predicts successful performance at a later time.

Variance: aggregate amount of variability in a set of scores.

Volunteer sample: sample of volunteers, rather than randomly assigned individuals.
Statistical Procedures

A

**Analysis of covariance**: an analysis of variance on dependent variable scores statistically adjusted for their relationship with a pre-test (called a covariate). When the covariate is related to the dependent variable, analysis of covariance is more powerful analysis than analysis of variance because it increases the likelihood of significant differences. Often abbreviated as ANCOVA.

**Analysis of variance, ANOVA**: statistical technique used to compare mean scores for two or more groups.

- **One-factor**: comparison of means on a dependent variable for one independent variable or one classification of subjects (association and descriptive studies).

- **Factorial**: comparison on a dependent variable for more than one independent variable or more than one classification of subjects. Interactions among independent variables are examined.

- **Multivariate**: tests for differences in the means on several dependent variables at one time, as opposed to a separate analysis of each dependent variable. Several independent variables can also be examined.

B

**Binomial test**: tests the hypothesis that a population proportion equals a specific value.

**Biserial correlation coefficient, \( r_b \)**: measures relationships when one variable is continuous and the second is dichotomous. It generally ranges in value from -1 to +1, although it may take on values below -1 or above +1.
**C**

**Canonical correlation**: produces two linear composites — one for dependent and one for independent variables — and examines the association between these in such a way that the correlation is a maximum value. If the two sets of variables have more than one significant relationship, more than one significant canonical correlation will be found.

**Chi-square (goodness of fit)**: compares frequencies actually obtained with expected frequencies. For example, compare the actual distribution of responses to an attitude scale (i.e., item 1 = agree; 2 = undecided; 3 = disagree) with the distribution that would have occurred if respondents had shown no preference for any position.

**Cluster analysis**: identifies items that “cluster” together as shown by item intercorrelation.

**Cohort analysis**: studies variables in one or more population cohorts.

**Contingency coefficient**: index of association between two nominal variables ranging from 0 to +1. The maximum value is restricted by the dimensionality of the cross-tabulation table formed by crossing the two variables.

**Correlation ratio**: index of association ranging from 0 to +1 that shows the strength of a curvilinear (nonlinear) relationship between two interval or ratio variables. The measure is asymmetric when a curvilinear relationship exists.

**Cramer’s V statistic**: index of association between two nominal variables ranging from 0 to +1.

**D**

**Discriminant analysis**: determines the combination of variables that maximizes differences among groups and the probability that an individual case belongs to each group.
**E**

**Eta squared:** (see Correlation ratio).

**F**

**Factor analysis:** examines patterns of association within sets of variables to determine whether a pattern of relationships exists so that the total number of variables can be reduced to a smaller number of factors.

**Friedman rank sum test:** nonparametric ANOVA that compares the ranks of matched subjects, or the same subjects are observed more than once.

**F-test:** tests the hypothesis that the variances of two populations are equal by determining the ratio of the two sample variances.

**G**

**Gamma, G or g:** index of association between two ordinal variables arranged in rank order ranging in value from -1 to +1.

**K**

**Kendall’s coefficient of concordance:** measures the degree of agreement between $m$ sets of $n$ ranks. For example, if 10 objects were ranked by 5 different judges, the coefficient of concordance is the degree of agreement among the judges.

**Kolmogrov-Smirnov one-sample test (goodness of fit test):** determines whether sample scores conform to a population characterized by a particular distribution, such as the normal or a chi-square distribution.

**Kruskal-Wallis analysis of variance:** nonparametric ANOVA that uses ranked scores to compare different groups.
Lambda, $\lambda$: measures the association between two bivariate distributions where both variables are interpreted as nominal variables. It simply reverses the role of two variables predicting $x$ from information about $y$. It ranges in value from 0 to +1.

Mann-Whitney U test: treats scores as ranks and rejects the null hypothesis if scores in one group are more likely to precede (or follow) scores in the other group. It is frequently used in place of the $t$-test when $t$-test assumptions cannot be met.

Median test: calculates the median for the total group and then compares the number in each group above and below this median,

Multiple correlation coefficient: measures the maximum relationship that can be obtained between a combination of several independent variables and a dependent variable.

Multiple regression: determines the association between a single criterion variable and a set of predictor variables. The strength of the association is the multiple correlation coefficient, $R$, which ranges in value from 0 to +1 and the square of this value, $R^2$. The relative contribution of each predictor to the criterion is reflected in regression coefficients or beta weights.

Omega squared, $w^2$: measures the association between a nominal variable and an interval/ratio variable ranging in value from 0 to +1. It is sometimes applied after a $t$-test or analysis of variance to provide an index of the percentage of variance accounted for in the criterion (interval/ratio) variable by the groups compared (nominal variable).
**P**

**Partial correlation coefficient**: measures the association between two variables where the relationship of each variable to a third variable has been removed statistically (“held constant” or “partialled out”).

**Path analysis**: variation of multiple regression analysis that identifies causal factors and estimates their strength in a causal chain.

**Pearson product-moment correlation, $r_x$**: measures the relationship between two variables when both are continuous and the relationship is rectilinear. It is most reliable when based on a large number of pairs of observations. The index ranges from $-1$ to $+1$.

**Phi coefficient, $\phi$**: measures the association between two dichotomous variables. It is a product-moment correlation applied to dichotomous data and ranges in value from $-1.00$ to $+1.00$.

**Point-biserial correlation coefficient, $r_{pb}$**: measures the relationship between a truly dichotomous and a continuous variable. It ranges in value from $-1$ to $+1$.

**Proportions, test of**: tests the hypothesis that two population proportions are equal.

**R**

**Rank biserial correlation coefficient, $r_{ib}$**: measures the association between a dichotomous and ordinal variable ranging in value from $-1$ to $+1$(sign is arbitrary).

**Rho, $\rho$**: (see Rank correlation coefficient).
**Sign test**: tests hypothesis that the difference between all pairs of scores from two related or dependent groups is zero. It only considers whether the sign of the difference between pairs is positive, negative, or zero. The null hypothesis is rejected when the number of pairs with positive differences is greater than the number of pairs with negative differences (or vice versa).

**Spearman rank order correlation** (rho or $r_s$): measures the association between two sets of ranked data. It ranges from +1 for a perfect match to -1 in ranks are exactly opposite.

**Tetrachoric correlation coefficient**, $r_{tet}$: measures the association between two dichotomous variables that have an underlying normal distribution. It ranges in value from −1 to +1.

**t-test**:  

*One-sample*: tests the hypothesis that a population mean equals a specified value.  

*Two-sample*: tests the hypothesis that two population means are equal.  

*Correlated/dependent samples*: tests the hypothesis that the population means of two related or dependent groups are equal. For example, test the same sample twice and use the $t$-test to examine the amount of change from one time to the next.

**Wilcoxon matched-pairs signed-rank test**: tests the hypothesis that differences in ranks between all pairs of scores is zero. Both the sign and the magnitude of the differences are considered with larger differences given more weight.
Yule’s Q: measures the association between two nominal variables ranging in value from $-1$ to $+1$. This special case of the gamma coefficient equals zero when one cell is empty.
References


Item Analysis for Criterion-Referenced Tests

by

Richard J. McCowan
Sheila C. McCowan