Houghton Mifflin Harcourt

HMH ASSESSMENTS

Glossary of Testing, Measurement, and Statistical Terms

Resource:
Joint Committee on the Standards for Educational and Psychological Testing of the AERA, APA, and NCME. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Glossary of Testing, Measurement, and Statistical Terms

Ability – A characteristic indicating the level of an individual on a particular trait or competence in a particular area. Often this term is used interchangeably with aptitude, although aptitude actually refers to one's potential to learn or to develop a proficiency in a particular area. For comparison see Aptitude.

Ability/Achievement Discrepancy – Ability/Achievement discrepancy models are procedures for comparing an individual's current academic performance to others of the same age or grade with the same ability score. The ability score could be based on predicted achievement, the general intellectual ability score, IQ score, or other ability score. Scores for each academic area are compared to the corresponding cognitive ability score in one of these areas to see if the individual is achieving at the level one would expect based on their ability level. **See also:** Predicted Achievement, Achievement Test, and Ability Testing.

Ability Profile – see Profile.

Ability Testing – The use of standardized tests to evaluate the current performance of a person in some defined domain of cognitive, psychomotor, or physical functioning.

Accommodation – A testing accommodation, refers to a change in the standard procedures for administering the assessment. Accommodations do not change the kind of achievement being measured. If chosen appropriately, an accommodation will neither provide too much or too little help to the student who receives it. Consequently, it seems reasonable to interpret that student's scores in the same ways the scores of all other students are interpreted. The student's scores can be included with all others in the group averages, and the various derived scores (e.g., grade equivalents and percentile ranks) should be reported without any flagging. From a measurement perspective, this "mainstream" processing approach seems reasonable because the changes made in test administration are not expected to change what is measured or to give the student an advantage over any other students. **See also:** Test Modification.

Accountability – The demand by a community for school districts to provide evidence that educational programs have led to measurable learning. "Accountability testing" is an attempt to sample what students have learned or how well teachers have taught, and/or the effectiveness of a school's principal's performance as an instructional leader. School budgets and personnel promotions, compensation, and awards may be affected. Most school districts make the results from this kind of assessment public; it can affect policy and public perception of the effectiveness of taxpayer-supported schools and be the basis for comparison among schools.

Achievement levels/Proficiency levels – Descriptions of an individual's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum, often labeled from "basic" to "advanced," that constitute broad ranges for classifying performance. The exact labeling of these categories may vary from one assessment or testing program to another. **See also:** Cut Score. ALDs describe what students should know and be able to do in a given content area and provide a measure for how well students' actual achievement matches the achievement desired on the assessment.

ALDs describe what students should know and be able to do in a given content area and provide a measure for how well students' actual achievement matches the achievement desired on the assessment.

Achievement Test - A test designed to measure the extent to which a person has acquired certain knowledge and/or skills that have been taught in school or as a part of some other planned instruction or training. The ***Woodcock-Johnson®* IV *Tests of Achievement*** and the ***Iowa Assessments*** are examples is an example of an achievement test.

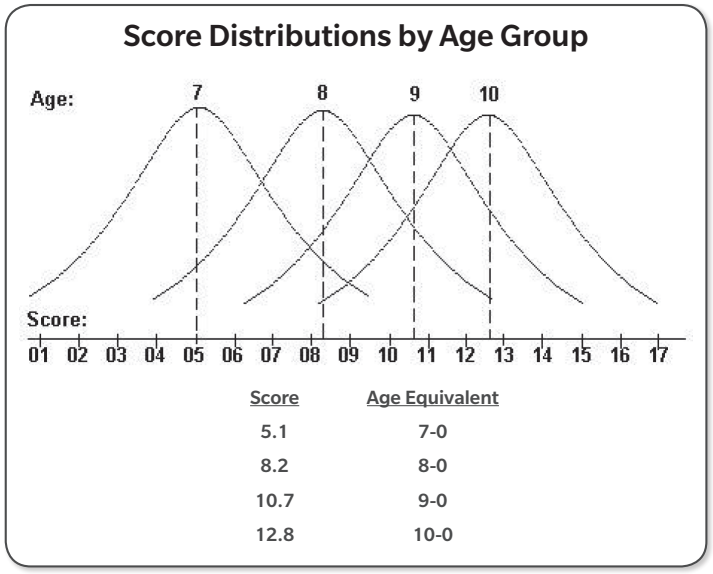
Adaption – An alteration to the administration of a standardized test or test item for a student in response to an individualized education plan or other requirement.

Adaptive Test (Computer Adaptive Test) – A form of individual or group testing in which items, or sets of items, are selected for administration to the test taker based primarily on the psychometric properties and content of the item, in relation to the test taker's responses to previous items. For example, as a student gets items correct, they will progressively be administered more difficult items. This creates an individualized testing experience where item selection is matched to the student's ability. ***Continuum Assessments™*** is an example of an adaptive test.

Adequate yearly progress (AYP) – A requirement of the No Child Left Behind Act (NCLB, 2001). This requirement states that all students in each state must meet or exceed the state-defined proficiency level by 2014 on state assessments. Each year, the minimum level of improvement that states, school districts, and schools must achieve is defined.

Age-Based Norms – Developed for the purpose of comparing a student's score with the scores obtained by other students at the same age on the same test. How much a student knows is determined by the student's standing or rank within the age reference group. For example, a norms table for 12 year-olds would provide information such as the percentage of 12 year-olds (based on a nationally representative sample) who scored at or below each obtainable score value on a particular test. Compare to Grade -Based Norms. **See also:** Norms, Standard Age Scores, Universal Scale Score, and Percentile Rank.

Age Equivalent – The chronological age in a defined population for which a given score is the median (middle) score. An examinee assigned an age equivalent of 7-5 indicates that he or she received the same score (raw score, scale score, standard score, etc.) as the average child who is seven years, five months old. **See Figure 1**. For comparison see Grade Equivalent.



Score	Age Equivalent
5.1	7-0
8.2	8-0
10.7	9-0
12.8	10-0

Figure 1: Age Equivalent Example

Age Percentile Rank – see Percentile Rank.

Age Stanine (AS) – see Stanine.

Aggregation - The compilation of the results of students for the purpose of reporting.

Alternate Forms – Two or more versions of a test that are considered interchangeable, in that they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. Alternate forms is a generic term. More specific terminology applies depending on the degree of (statistical) similarity between the test forms - parallel forms, equivalent forms, and comparable forms - where parallel forms refers to the situation in which the test forms have the highest degree of similarity to each other.

Analytic Scoring – A method of scoring in which a student's performance is evaluated for specific traits or dimensions. Each critical dimension is judged and scored separately, and the resultant values are combined for an overall score. In some instances, scores on the separate dimensions may also be used in interpreting performance. Contrast to Holistic Scoring.

Anchor Papers – A set of papers showing student's responses to constructed-response items, selected in a score range-finding activity, that represent exemplar responses of each possible score point. Anchor papers are used during training sessions to guide raters on how to score open-ended items. **See also:** Constructed-Response Item.

Ancillary Examiner – An ancillary examiner is preferably a native speaker of the target language who is trained in the administration of the test. The major responsibility of the ancillary examiner is the precise administration of test items in a foreign language and the observation of test behavior. These examiners work under the supervision of a primary examiner, whose responsibility it is to calculate the derived scores and interpret the test results.

Answer Choice – All options available for a student to select from in a multiple-choice item.

Aptitude Test – A test designed to measure the ability of a person to develop skill or acquire knowledge. Any test used to assess a person's "readiness" to learn or potential for future success in a particular area if the appropriate education or training is provided (e.g., musical aptitude test). An IQ test is a well-known example of an aptitude test for measuring general academic ability. **See also:** Readiness Test and IQ Test.

Arithmetic Mean – The average of a set of scores obtained by adding all scores in the set and dividing by the total number of scores. For example, the mean of the set {85, 76, 97, and 66} is 81. The arithmetic mean is commonly referred to as simply the mean. **See Figure 2. See also:** Average and Central Tendency.

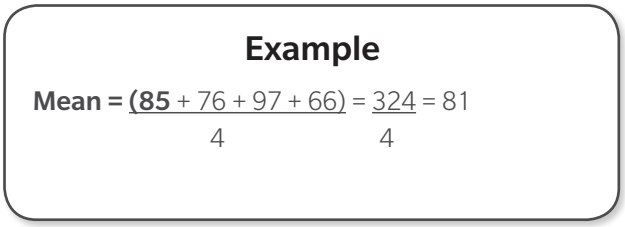


Figure 2: Arithmetic Mean Example

Assessment – Any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs. Assessment aids educational decision-making by securing valid and reliable information from multiple sources of evidence (combinations of performance, products, exhibits, discourse, tests, etc.) and integrating this information in order to judge whether students have learned what is expected. **See also:** Formative Assessment, Summative Assessment.

Augment - Altering the content of an existing assessment by removing, adding, or changing items.

Average – Any statistic indicating the central tendency or most typical score of a set of scores – such as the arithmetic mean, median, and mode. The term average, without qualifications as to type, typically, refers to the arithmetic mean. **See also:** Central Tendency, Arithmetic Mean, Median, and Mode.

Basal – For individually administered tests, the point on test, associated with a given level of functioning or skill, for which an examiner is confident, that all items prior to that item would be answered correctly (considered too easy). The items below this point, although not administered to the individual student, are afforded full credit. Contrast to Ceiling. Basal and ceiling rules act to enhance the efficiency of the test administration process by administrating only the range of items required to obtain an accurate estimate of the individual's ability.

Basic Interpersonal Communication Skills (BICS) – BICS is defined as language proficiency in everyday communication. This language is acquired naturally, without formal schooling. Conversing with family members and friends using skills in pronunciation, basic vocabulary, and grammar is an example of BICS.

Battery – A set of tests, generally standardized on the same population, and designed to be administered as a unit. The scores on the several tests usually are scaled so that they can be readily compared or used in combination for decision-making. In other words, scores can be computed and interpreted for each individual test as well as the entire battery. The ***Woodcock-Johnson* IV *Tests of Early Cognitive and Academic Development*** or any of the main ***Woodcock-Johnson* IV** sub-tests, such as the ***Tests of Oral Language***, fall into this category. **See also:** Complete Battery.

Benchmark – A statement providing a description of what students are expected to learn at various developmental levels or points in time (e.g., the 3rd grade) and is used to indicate a student's progress toward meeting specific content standards. **See also:** Content Standard.

Bias – In a statistical context, bias refers to any source of systematic error in the measurement of a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct-irrelevant compone nts of test scores that differentially affect the performance of different groups of test takers (e.g., gender, ethnic, or age, etc.). Test developers try to reduce bias by conducting item fairness reviews and various item analyses, detecting potential are as of concern, and either removing or revising these types of test items prior to the development of the final operational form of the test. **See also:** Systematic Error, Item Analysis, Fairness/Sensitivity Review, and Construct.

Building Identification (ID) Sheet – A machine-scannable form used to identify, for reporting purposes, the name of the building or school in which a particular set of answer documents was administered.

Cattell Horn-Carroll (CHC) theory – Theory of cognitive abilities, which stems from the psychometric factor-analytic work of Raymond Cattell (1941, 1943, 1950), John Horn (1988, 1991), and John Carroll (1993, 1998). CHC has both broad and narrow cognitive abilities, including oral language, reading, mathematics, writing, and academic domain-specific aptitudes; and academic knowledge (Schneider & McGrew, 2012).

Cattell Horn-Carroll (CHC) theory – Theory of cognitive abilities, which stems from the psychometric factor-analytic work of Raymond Cattell (1941, 1943, 1950), John Horn (1988, 1991), and John Carroll (1993, 1998). CHC has both broad and narrow cognitive abilities, including oral language, reading, mathematics, writing, and academic domain-specific aptitudes; and academic knowledge (Schneider & McGrew, 2012).

Ceiling – The upper limit of ability that a test can effectively measure or for which reliable discriminations can be made. Individual or groups scoring at or near the highest possible score are said to have "reached the ceiling" of the test (i.e., a ceiling effect), and should, if possible, be administered the next higher level of the test in order to obtain a more accurate estimate of their ability. For individually administered tests, the ceiling refers to the point during administration, after which, all other items will no longer be answered correctly (considered too difficult), and results in the examiner stopping the administration of the test. Contrast to Floor and Basal.

Central Tendency – A term that refers to the center or middle of a score distribution, typically represented by some type of average. The mean and median are measures of central tendency. **See also:** Average, Arithmetic Mean, and Median.

Chance Success Level – In a testing context, specifically for selected–response items, the chance success level is the proportion of correct answers that could occur if all examinees were to randomly guess at the answer. For example, for a five-response option multiplechoice item, each examinee has a 1 in 5 chance of getting the item correct by simply guessing. The chance success level for this item would be 20%. Knowing the chance success level for each item, one can estimate the chance level raw scores for the test – the range of scores that could have been achieved by randomly marking answers to each test item. For example, if a 40-item multiple-choice test is composed entirely of five-response option items, it would be possible that up to 20% of the points could have been achieved by chance alone (40 x .20 = 8). The chance level raw scores for this test are from a raw score of 0 to a raw score of 8. On some tests, these chance raw scores are referred to as Targeted Scores.

Checklist – In assessment, a list of characteristics or behaviors, used by an examiner as a guide for evaluating an individual or group's performance, by noting the presence or absence of each assessed cha racteristic or behavior. For example, a psychologist observing children at play might assess the degree of cooperative behavior by determining which of a set of behaviors or characteristic are displayed (e.g., sharing toys, taking turns, types of communication, etc.).

Classical Test Theory – A psychometric theory that represents the factors influencing an individual's observed test score. The model postulates that an individual's observed score is the sum of a true score component plus an independent error component (i.e., any unsystematic, or random, influence of an individual during testing, such as fatigue, practice effects, individual's mood, or effects of environment). Many standard procedures for test construction and for the evaluation of a test's reliability and validity are based on the assumptions of this model. **See also:** True Score, Error of Measurement, Reliability, Validity.

Classroom-Based Assessment – An assessment developed, administered, and scored by a teacher or group of teachers to evaluate a student's or classroom's performance on a topic.

Coefficient Alpha (Cronbach's Alpha) – see Reliability Coefficient, Reliability, and Internal Consistency.

Cognitive Academic Language Proficiency (CALP) – CALP is defined as language proficiency in academic situations, or those aspects of language proficiency that emerge from formal schooling.

Cognitive Assessment – The process of systematically gathering test scores and related data in order to make judgements about an individual's ability to perform various mental activities involved in the processing, acquisition, retention, conceptualization, and organization of sensory, perceptual, verbal, spatial, and psychomotor information. The ***Woodcock-Johnson* IV *Tests of Cognitive Abilities*** battery falls into this category.

Cohort – A group of individuals with a common demographic (e.g., age or class membership) whose progress is followed by obtaining measurements (e.g., test scores on the same developmental scale) at different points in time (e.g., 3rd, 5th, and 8th grades, etc.).

Complete Battery – A collection of tests that measure a broad range of skills. For example, the Complete Battery of the ***Iowa Assessments***™ measure Reading, Written Expression, Mathematics, Science, Social Studies, Vocabulary, Language, Spelling, Capitalization, Punctuation, Computation, Word Analysis, and Listening (depending on the Level). ***Woodcock-Johnson IV*** is also a complete battery when the T***ests of Achievement, Tests of Cognitive Abilities***, and ***Tests of Oral Language*** are administered. **See also:** Battery.

Completion Rates – The percent of test takers completing the entire test (e.g., 100% of the items) or the percent of test takers completing a specific percent of the number of items (e.g., 75% of the items). Examining completion rates can be useful for determining the perceived speededness and/or difficulty of the test, both of which are characterized by low completion rates. **See also:** Speededness.

Composite Score – A score that is derived by combining one or more scores according to a specified formula. This is typically accomplished by averaging or summing the contributing scores which are often weighted according to their relative importance. **See Figure 3**. **See also:** Average.

Examples:

Reading Total = ((Reading Subtest #1 Score) + (Reading Subtest #2 Score)) / 2

Math Total = (Math Subtest #1 Score) + 2 (Math Subtest #2 Score)

Figure 3: Composite Score Example

Confidence Band – see Confidence Interval.

Confidence Interval – An interval between two values on a score scale within which, with specified probability, a score or parameter of interest lies. An individual's test score provides a good point estimate of the student's ability in a specific area. However, this estimate, as in any measurement process, contains some degree of error (either for or against the student's favor). A confidence interval provides a range of values around the estimate to indicate the how accurate or precise the estimate is likely to be. The confidence level associated with the score interval, usually 68%, 80%, or 90%, indicates the percentage of times, given repeated sampling, that the interval will contain the student's true score. For example, if one chose a 68% confidence interval and the test were administered 100 times (in theory, assuming no practice effects or change to the student), 68 out of 100 times, the true score for the student would fall within the given score interval. Confidence intervals are constructed using the student's observed score and information about the test's standard error of measurement (SEM). **See Figure 4**. **See also:** True Score and Standard Error of Measurement.

Confidence Interval Lower Limit = Student's Score - (Z)(SEM)
Confidence Interval Upper Limit = Student's Score + (Z)(SEM)
Where: Z is a number that varies depending on the confidence level:

68% confidence level:	Z = 1.00
80% confidence level:	Z = 1.28
90% confidence level:	Z = 1.65

Example:

Student's Score = 10
Test's Standard Erro r of Measurement = 3.0

68% confidence interva	= 10 - (1)(3) to 10 + (1)(3)
	= (10 - 3) to (10 + 3)
	= 7 to 13
80% confidence interval	= 10 - (1.28)(3) to 10 + (1.28)(3)
	= (10 - 3.84) to (10 + 3.84)
	= 6.16 to 13.84
	˜ 6 to 14
90% confidence interval	= 10 - (1.65)(3) to 10 + (1.65)(3)
	= (10 - 4.95) to (10 + 4.95)
	= 5.05 to 14.95
	˜ 5 to 15

Figure 4: Confidence Interval Example

Co-norm – A term indicating that two or more tests have been concurrently normed or standardized at the same time with the same group of test takers (norm group). The ***Iowa Assessments*** are co-normed with ***CogAT***® (***Cognitive Abilities Test***™). **See also:** Standardization, Standardization Sample, and Norms.

Consequential Validity – see Validity.

Construct – The concept or the characteristic that a test is designed to measure, but is not directly observable (referred to as a latent variable). A construct is a theoretical concept or trait inferred from multiple evidences and used to explain observable behavior patterns. In psychological testing, a characteristic that is considered to vary across individuals, such as extroversion, visual-spatial ability, creativity, etc. **See also:** Validity-Construct Validity.

Constructed-Response Item – An item format, designed in a specific way that requires examinees to create their own responses, which can be expressed in various forms, (e.g., write a paragraph, create a table/graph, formulate a calculation) rather than choose from a supplied set of answers (e.g., multiple-choice or forced-choice items). Types of constructedresponse items include Short-answer and Extended Response items. These items are a form of performance assessment and are commonly referred to as open-ended items. Contrast to Multiple-Choice Item. **See also:** Short-Answer Item, Extended-Response Item, and Performance Assessment.

Construct-Related Validity – see Validity.

Content Area – This term is used interchangeably with "subject area", for example, Language is a content area. **See also:** Subject Area.

Content Classification (Content Domain) – A set of organized categories by which items are classified that detail the specific knowledge, thinking skills, abilities, or other characteristics to be measured by a test.

Content-Related Validity – see Validity.

Content Standard – A statement of a broad goal describing expectations for students in a subject area at a particular grade or at the completion of a level of schooling. Content standards are the goals of the curriculum. The standards are typically the basis or specification on which the test items are written.

Correlation – The degree of relationship (or strength of association) between two sets of scores, based on the same set of individuals. The relationship can be positively correlated (e.g., students scoring high on one test also tend to score high on the other), negatively correlated (e.g., students scoring low on one test tend to score high on the other), or zero correlated (e.g., lack of any relationship be tween the scores). Correlation simply refers to the strength of the relationship existing between two sets of values and does not necessarily imply that one influenced the other or caused the other to happen. The most commonly used statistic for measuring correlation is *Pearson's product moment r* (Pearson's correlation coefficient). **See also:** Correlation Coefficient.

Correlation Coefficient – A statistic used to measure the strength of a relationship between two sets of scores, based on the same set of individuals. The correlation coefficient, denoted r, ranges in value from -1 to +1. A correlation of +1 or -1 indicates a perfect (positive or negative) relationship, while a correlation of 0 (zero) indicates the complete absence of a relationship. The closer the value of the correlation coefficient is either to -1 or +1, the stronger the relationship. Correlation coefficients are used in a variety of circumstances including estimating some types of test reliability (e.g., test-re-test and splithalves reliability) and validity (e.g., criterion-related and construct-related validity). The most commonly used statistic for measuring correlation is *Pearson's product moment r* (Pearson's correlation coefficient). **See also:** Correlation, Test-Re-Test Reliability, Split-half Reliability Coefficient, and Validity.

Criterion – A standard, guideline, or rule by which a judgement or decision may be based. For example, an outlined objective by which a student's response or performance is judged.

Criterion-Referenced Interpretation – The comparison of a student's score to specific criteria or standard of performance rather than in relation to the overall ranking of that student's score with respect to his or her age or grade peers (norm group). The decision of whether or not a student has mastered a skill or demonstrated a minimal acceptable performance involves a criterion-referenced interpretation. In classroom-based testing, a teacher may determine the score needed for mastery or passing based on a specified percent correct score. At times, criterion-referenced interpretations are also made about tests that provide norm-referenced information. Criterion-referenced interpretations in standardized tests are typically determined by setting standards of performance and/or by using cut scores to determine the various levels of proficiency. See other types of Criterion-Referenced Tests: Classroom-Based Assessment, Curriculum-Based Assessment, and Mastery Test. **See also:** Performance Standard, Percent Correct, Cut Score, Standard Setting. Contrast to Norm-Referenced Interpretation.

Criterion-Referenced Test (CRT) – A test designed to measure a student's performance as compared to an expected level of mastery, educational objective, or standard. The type of scores resulting from this type of test provide information on what a student knows or can do in with respect to a given content area as opposed to a score indicating how that student ranks among his or her age or grade peers (norm group). For contrast see Norm-Referenced Test (NRT). **See also:** Percent Correct, Cut Score, Criterion-Referenced Interpretation, Standard Setting.

Criterion-Related Validity – see Validity.

Critical Thinking Skills – Using higher-order thinking skills (abstract reasoning and problem solving) to lead to an understanding of more complex tasks, problems, or ideas. Critical thinking skills require more than simple recall in arriving at solutions and include: interpretation, analyzing, comparing, classifying, applying logic, and making inferences.

Cronbach's Alpha (Coefficient Alpha) – see Reliability Coefficient, Reliability and Internal Consistency.

Curriculum Alignment – The degree to which the curriculum objectives match a testing program's evaluation instruments (e.g., test's content specifications).

Curriculum-Based Assessment – A type of assessment used to evaluate a student's progress that is closely aligned to instructional materials.

Cut Score - (Cutoff Score) – A specified point on a score scale such that scores at or above that point are interpreted or acted upon differently from scores below that point. For example, a score designated as the minimum level of performance needed to pass a competency test. One or more cut scores can be set for a test which results in dividing the score range into various proficiency level ranges. Methods for setting cut scores vary. **See also:** Achievement levels/Proficiency levels, Performance Standard, and Standard Setting.

Decile – One of nine points (scores) that divides a distribution of scores into ten equal groups, each containing one -tenth (10%) of the data. Deciles are special cases of percentiles - every tenth percentile. The second decile is the 20th percentile; the ninth decile is the 90th percentile; etc. **See also:** Percentile.

Derived Score – A score to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to standard scores, percentile ranks, grade equivalents, stanines, etc. **See also:** Raw Score, Standard Score, Scale Score, Percentile Rank, Grade Equivalent, Stanine.

Developmental Scores – A type of score that is used to show an individual's position along a developmental continuum. Developmental Scores allow comparisons to be made to a series of reference groups that differ systematically and developmentally in average achievement, usually age or grade groups. Grade Equivalents, Age Equivalents, Developmental Standard Scores, and Scale Scores are all types of developmental scores. See Developmental Standard Score, Scale Score, Grade Equivalent, Age Equivalent.

Developmental Standard Score (DSS or SS) – The Developmental Standard Score provides a continuous growth scale of achievement from kindergarten through grade 12 for such tests as the ***Iowa Assessments***. The DSS is the standard score that is used for entry to the grade norms tables to obtain such derived scores as the Grade Percentile Rank (PR) and Grade Equivalent (GE) for each test and composite score. The Grade Percentile Ranks can then be converted to other derived scores such as the Grade Stanine by the use of another set of conversion tables. **See also:** Standard Score, Grade-Based Norms, Derived Score, Grade Equivalent, Percentile Rank, and Stanine.

Diagnostic Test – An intensive, in-depth evaluation process with a relatively detailed and narrow coverage of a specific area, generally prompted by a perceived problem and administered in order to determine a student's current level of functioning. A diagnostic test is primarily used to identify specific areas of strength or weakness, locate learning difficulties, or to identify special needs; determine the nature of any deficiencies; and if possible, to identify their cause. The results of a diagnostic test are used to prescribe a solution to meet the learning needs of the student through either regular or remedial classroom instruction.

Difficulty Index – see Item Difficulty, P-value.

Discrimination Index – see Item Discrimination, Point Biserial.

Distractor – An incorrect response option for a multiple-choice, multiple-select, or matching item, but one that is plausible because it is aligned with common conceptual or computational errors made by test takers. That is, it is a response that might be deemed likely for those with less content knowledge, skill, or ability in the assessed area. The distractors must be clearly incorrect or not the best option. If the distractor seems so implausible that no one will select the distractor, it does not add value to the item.

Distribution – A tabulation of scores (ordered either in ascending or descending value) showing the number of individuals in a group obtaining each score or contained within a specified fixed range of scores (score interval). Also commonly referred to as a Frequency Distribution. **See Figure 5**. **See also:** Frequency.

Examples:

Score	Frequency
4	1
10	2
12	5
15	4
16	5
17	6
20	5
23	4
25	2
30	1

Score Interval	Frequency
0 – 5	1
6 – 10	2
11 – 15	9
16 – 20	16
21 – 25	6
26 – 30	1

Figure 5: Distribution Example

Empirical Norms Dates – The empirical norms date is the time interval (or week) that contains the median test administration date for the norming sample (i.e., standardization sample). Most testing programs allow score reporting of the empirical norms for a given period of time (e.g., a period of five-weeks -2 weeks prior to and 2 weeks after the median test date). Some testing programs have developed interpolated norms for those schools that test outside of the empirical norms dates, but still desire normative information. Schools that have academic calendars starting earlier or later than is customary (e.g. other than late August/early September) should establish their testing schedules based on the number of instructional days into the academic year. For example, if the empirical norms date represents the week in which the thirtieth day of instruction falls for a school with a regular academic calendar, these schools should test approximately thirty days into instruction as well, based on their academic calendar. For more specific information on the empirical norms dates, consult the interpretative guide and technical manual for the test administered. **See also:** Norms, Standardization Sample, and Interpolated Norms.

Equating – A technical or statistical procedure or process by which scores from one or more alternate forms of a test are converted to a common scale or metric for the purposes of comparability and determination of score equivalence. The goal is to establish comparable scores on different versions of a test, allowing them to be used interchangeably. There are many methods available to equate or "link" the test scores from one test to the test scores on another test. It is an important aspect of establishing and maintaining the technical quality of a testing program. After equating, educators can interpret performance on one test form as having the same substantive meaning compared to the equated score on the other test form.

Equivalent Forms – see Alternate Forms.

Error of Measurement – The difference be tween an individual's observed score (actual score received) and their underlying true score (a theoretical construct referring to an error-free assessment of the person's ability). **See also:** Standard Error of Measurement and True Score.

Expected Growth – The amount of change in test scores that occurs over a specified amount of time, based on individuals that share certain characteristics such as age or grade level. There are a variety of methods available for computing expected growth, such as normative scores, gain score, trajectory, projection, student growth percentiles, etc. For example, by plotting normative data, scale scores associated with various percentile ranks (10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}), across ages or grades for a given subject area, the expected growth can be determined for individuals at various functional levels. Based on this information, it is possible to determine if a student is growing faster than or slower than expected in relationship to other students in his/her grade or age group. **See also:** Norms, Percentile Rank, Scale Score, and Standard Score.

Extended-Response Item – An item test format that is a type of constructed-response question that requires students to formulate an extended essay response. **See also:** Constructed-Response Item.

Extended-Time – The instructions for some tests allow certain students to be administered the test under extended-time conditions. That is, the time limi ts are extended beyond that allowed by a regular administration of the test to reduce the effect of slow work rate on a student's test performance. Certain tests may provide a separate set of norm conversions developed using a reference group who were administered the test under the same extendedtime conditions. For more specific information, consult the interpretative guide and technical manual for the test administered. **See also:** Accommodation and Norms.

Face Validity – see Validity.

Factor – Any variable, real or hypothetical, which is an aspect of a construct. An unobserved variable, that is, not directly measurable, that can be thought of as a label that characterizes the responses to a related group of observable variables. For example, creativity is a quality that is not directly measurable but is inferred from observable activities such as drawing, writing, role-playing, etc. In measurement theory, a factor is a statistical dimension defined by a factor analysis. **See also:** Construct, Factor Analysis, and Factor Scores.

Factor Analysis – Any of several statistical methods of describing the interrelationships of a set of observed variables (e.g., items or score totals) by statistically deriving new variables, called factors, that are fewer in number than the original set of variables. Factor analysis techniques are used to identify the underlying dimensions within a set of data and to provide evidence that the underlying structure is consistent with the theoretical constructs intended by the test's design. **See also:** Factor and Factor Scores.

Factor Scores – A score that is an estimate of how an examinee would have scored on a factor, had it been possible to measure the factor directly. Computationally, factor scores are a linear combination of item scores and factor score coefficients, which result from conducting a factor analysis. Once computed, factor scores can be used for subsequent analyses for establishing the validity of the factor structure. **See also:** Factor, Factor Analysis, and Validity - Construct-Related Validity.

Fairness/Sensitivity Review – A review of the items on a test in order to determine if the items are assessing all test takers in an equitable manner. The process typically involves a panel consisting of both male and female professional educators representing various ethnic and racial groups who are asked to use their personal and professional judgement to critique the items for offensive or stereotypical content.

Fall Norms –Taking into account the date of testing is an important factor in normreferenced interpretations. Students testing in the fall of a given grade level have not been given the same amount of instruction as those tested in the spring. Because of this, norm tables (specifically National Percentile Ranks) are provided for different times of the year. Fall norms are typically provided for those students testing from August to November, however this time frame may vary depending on the test that was administered. For more specific information, consult the inte rpretative guide and technical manual for the test administered. **See also:** Norms, Norm-Referenced Interpretation, and Percentile Rank.

Field Test – A test administration used to check the adequacy of testing procedures, generally including the test administration directions, test responding, test scoring, and test reporting. A field test is generally more extensive than a pilot test. See Pilot Test.

Floor – The lower limit of ability that a test can effectively measure or for which reliable discriminations can be made. Individual or groups scoring at or near the lowest possible score are said to have "reached the floor" of the test, and should, if possible, be administered the next lower level of the test in order to obtain a more accurate estimate of their ability. Contrast to Ceiling.

Formative Assessment - The use of tools and techniques to measure student understanding prior to the completion of teacher instruction. The information collected through formative assessment is used to revise or refine future lesson plans to addresses areas of deficiency.

Frequency – The number of times that a certain value or range of values (score interval)occurs in a distribution of scores. See Distribution for an example.

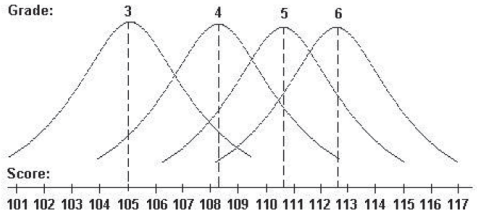
Frequency Distribution (FD) – see Frequency and Distribution.

Grade-Based Norms – Developed for the purpose of comparing a student's score with the scores obtained by other students in the same grade on the same test. How much a student knows is determined by the student's standing or rank within the grade level score distribution. For example, a norms table for the 6th grade would provide information such as the percentage of 6th grade students (based on a nationally representative sample) who scored at or below each obtainable score value on a particular test. Compare to Age-Based Norms. **See also:** Norms, Standard Score, Grade Equivalent, and Percentile Rank.

Grade/Class Identification (ID) Sheet – A machine-scannable form used to identify, for reporting purposes, the name of the class and grade level in which a particular set of answer documents was administered.

Grade Equivalent – The grade equivalent is a number that describes a student's location on an achievement continuum. Grade Equivalents are expressed in terms of grade and months into grade, assuming a 10-month school year (e.g., 8.4 means after 4 months of instruction in the 8th grade). The Grade Equivalent corresponding to a given score on any test indicates the grade level at which the typical student obtains this score. Because of this, Grade Equivalents are not based on an equal interval scale, and therefore cannot be added, subtracted, or averaged across test levels the wa y other scores can (scale scores or standard scores). **See Figure 6.** For comparison see Age Equivalent.

Score Distributions by Grade Group



Standardization sample was tested in October (2 months into instruction).

Score	Grade Equivelent
105.0	3.2
108.3	4.2
110.6	5.2
112.8	6.2

Interpretations/Cautions of Grade Equivalent Scores:

If a sixth-grade student obtains a GE of 8.4 on the Vocabulary Test, his or her score is like one a typical student finishing the fourth month of eighth grade would likely get on the Vocabulary Test.

Grade Equivalents are suited for estimating a student's developmental status or year-to-year growth. These scores are not suited for identifying a student's standing within a group or for diagnosing areas of relative strength or weakness. Grade Equivalents do not provide a prescription for grade placement. For example, a fourth-grade student earning a GE of 6.2 on a fourth-grade reading test obviously demonstrates that his/her developmental level in reading is high relative to his/her fourth-grade peers. However, the test results supply no information about how well he/she would handle the material normally read by students in the early months of sixth grade. It, therefore, cannot be concluded that he/she is ready for placement in the sixth grade.

Figure 6: Grade Equivalent Example

Grade Stanine (GS) – see Stanine.

Growth – For an individual or group, the amount of change or difference between two tests scores, measured on the same score scale, resulting from the same test (or equated forms of the test) administered at two different points in time. Growth can be calculated by determining differences in the developmental scores (Developmental Standard Scores, Scale Scores, Grade Equivalents, etc.) from one test administration to another. **See also:** Expected Growth, Developmental Scores, Developmental Standard Score, Scale Score, Grade Equivalent, Alternate Forms, and Equating.

High Stakes Testing – A test that has important, direct consequences for the examinees, programs, or institutions involved in the testing. One example of such a test is an instrument that has minimum competency cutscores in which decisions to provide sanctions to schools or to withhold students from promotion to higher grades are made on the basis of student test results.

Holistic Scoring – A scoring method for obtaining a score on a test, or a test item, based on a judgement of overall performance using specified criteria or scoring rubrics. The procedure yields a single score representing the overall effectiveness of the individual's performance on the task (e.g., a constructed-response item) as opposed to analytical scoring methods that evaluate each critical dimension separately and then combine the resultant subscores for an overall score. Contrast to Analytic Scoring. **See also:** Rubric and Constructed-Response Item.

Individualized education program (IEP) - A document for a child that specifies which test accommodations that a student is eligible for. An IEP is developed by a team which typically includes the student's parents, teacher, and school psychologist.Infit - Fit statistics are used to determine how well a particular item or person is 'fitting' with the IRT model being implemented. The infit statistic is sensitive to the moderate persons or items in the dataset. This statistic will range from zero to infinity, with the target value being 1.0. **See also:** Outfit.

Intelligence Test – A psychological or educational test designed to measure an individual's level of cognitive functioning (verbal reasoning, abstract reasoning, memory, etc.), in accord with some recognized theory of intelligence.

Intercorrelations – A matrix of correlation coefficients, calculated between two or more sets of scores for the same set of individuals. **See also:** Correlation and Correlation Coefficient.

Internal Consistency – The degree of relationship among the items of a test. **See also:** Reliability.

Interpolated Norms – Typically norms are developed from data gathered at one or two points during the school year (a empirical testing window in the spring and/or fall) and the normative information (i.e., National Percentile Ranks) for the time interval between these points is approximated by interpolation. Interpolated norms are developed to provide a better estimate of normative information throughout the school year for those schools that desire the flexibility of testing at any time during the year or outside the empirical testing window. **See also:** Interpolation, Norms, Percentile Rank, Empirical Norms Dates, Quartermonth Norms.

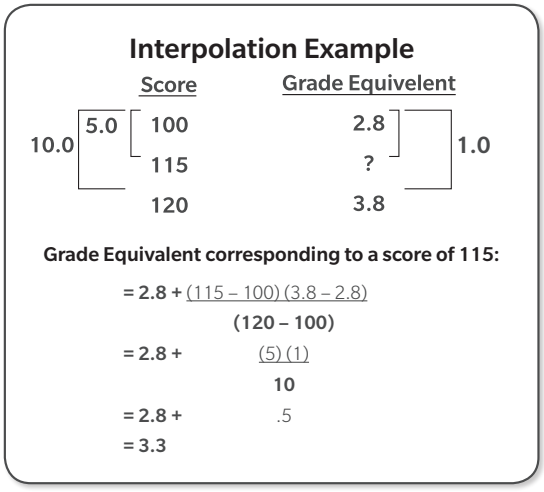


Figure 7: Interpolation Example

Interpolation – Any process of estimating intermediate score values between two known score points. Linear interpolation methods assume that the intermediate score values are equally spaced (i.e., assumes an equal interval scale) within the score interval defined by the known score points. **See Figure 7. See also:** Interpolated Norms.

Interquartile Range – A computed value indicating the distance between the lower quartile and upper quartile. Computed as the upper quartile minus the lower quartile. When used as a measure of the spread or variability of scores in a distribution, it is less sensitive to the possible presence of outliers than either the Standard Deviation or the Range. **See also:** Quartile, Standard Deviation, and Range.

Inventory – A questionnaire or checklist, usually in the form of a self-report, which elicits information about an individual's personal opinions, interests, attitudes, preferences, personal characteristics, motivations, and typical reactions to situations and problems. For example, a career interest inventory could be given to students for the purpose of eliciting information about interests and likes/dislikes with regard to certain activities in order to suggest or direct these individuals into careers that would be most suited for them. **See also:** Checklist.

IQ Test – An intelligence test administered to measure an individual's IQ (Intelligence Quotient). IQ scores were originally expressed as the ratio of an examinee's mental age to his/her chronological age, although over the years, that formula has been replaced with the concept of the deviation IQ (e.g., a standard score with a mean of 100 and a standard deviation of 16). **See also:** Intelligence Test, Standard Score.

Item – A general term referring to a single statement, question, exercise, problem, or task on a test or evaluative instrument for which the test taker is to select or construct a response, or to perform a task. Includes all the elements of an item as a collective unit: the stem, response options, prompt, stimulus, etc. **See also:** Stem, Prompt(Stimulus), Foil, Multiple-Choice Item, and Constructed-Response Item.

Item Analysis – The process of studying examinee responses to single test questions in order to determine the quality of each item with respect to certain characteristics such as item difficulty, item discrimination, and correlation with an external criterion, etc. **See also:** Item Difficulty, Item Discrimination, P-value, Point Biserial, and Correlation.

Item Characteristic Curve - The graph of an item based on data collected which shows a student's ability level versus the probability of answering an item correctly. Up to 3 parameters may be represented on the graph and they include: (a) the slope, or discrimination of the item, (b) the difficulty of the item, and (c) the guessing parameter of the item.

Item Difficulty – The extent to which an item presents itself as a challenge to a specific group of test takers. In Classical Test Theory, the item mean can be used as an indication of an item's difficulty. The higher the item mean, the easier the item is for the group; the lower the item mean, the more difficult the item is for the group. For most multiple-choice items or other item types in which the item is only worth 1 point, the item mean is equivalent tothe proportion of examinees answering the item correctly. **See also:** P-value, Item, Multiple-Choice Item, Constructed-Response Item, and Classical Test Theory.

Item Discrimination – The extent to which an item on a test differentiates between those examinees possessing much of the trait or skill being measured (high scorers) from those possessing little of the trait or skill being measured (low scorers). In Classical Test Theory, item discrimination indices generally range from -1.0 (little or no differentiation) to +1.0 (high differentiation), with higher values, in general, more desirable, and those with negative values considered extremely undesirable. Methods of determining an item's discriminating power can vary. **See also:** Point Biserial and Classical Test Theory.

Item Norms – Norms, specifically a set of item p-values (i.e., item difficulties), that are developed from distributions of student item-level scores, using a nationally representative sample of students. Item norms indicate the difficulty of each item as determined by data gathered from the standardization sample, by which future comparisons are made (e.g., classroom, building, district, or state item-level score averages). **See also:** Norms, Pvalue, Item Difficulty, and Standardization Sample.

Item Response Theory - A mathematical model of the relationship between performance on a test item and the test taker's level of performance on a scale of the ability, trait, or proficiency being measured, usually denoted as theta. Parameters are determined for both items (i.e., an item's discrimination, difficulty, and/or guessing parameter) and test takers (i.e., theta) in such a way as to probabilistically determine the response of any examinee to any item. Commonly referred to as a strong true -score theory or latent trait model and serves as the basis for a number of statistical models for analyzing items and test data. There are three common IRT models used in assessment - the three-parameter logistic model, which estimates three values for items including the a-parameter (dissemination parameter), b-parameter (difficulty parameter), and c-parameter (probability low-ability examinees are guessing). The two-parameter logistic model estimates values for the a-parameter and b-parameter and one-parameter logistic model, or Rasch, estimates just the b-parameter. **See also:** Item, Ability, True Score, Item Difficulty, Item Discrimination.

Item Tryout - see Pilot Test.

K-R 20 (Kuder-Richardson Formula 20) - A formula used to measure the internal consistency (i.e., reliability) of a set of dichotomously scored items (i.e., scored either "1" or "0"), based on a single administration of the test. Calculated using information about the pvalue for each item, the variance of the total test, and the total number of test items. **See also:** Reliability, Internal Consistency, P-value, and Variance.

Large City Norms - Norms (specifically Percentile Ranks) that are developed from distributions of student test scores using a representative sample of students residing in large cities only (e.g., populations of 250,000 and over). Large City Norms are used to indicate the status or relative rank of a student's score compared to other students in the nation also residing in highly populated areas. Educational programs and populations in large cities have unique characteristics that affect test performance, and therefore, separate norm tables are provided. See Norms, Percentile Rank, Distribution, and Standardization Sample.

Lexile® Framework - A tool for connecting the reading ability of a student (based on a student's lexile score) with reading materials (list of books) that have a suitable level of difficulty for the student - that is, reading material that is challenging, but not frustrating. Lexile scores for both students and reading texts are reported as a number followed by a capital "L" (Lexile) and range in value from below 0 to above 2000L. In theory, a student with a lexile measure of 600L who is given a 600L reading text, is expected to have a 75% comprehension rate. The Lexile Framework is available with the reporting of the *Gates- MacGinitie Reading Tests®* and the ***Iowa Assessments*** (*Iowa Tests of Basic Skills™* and *Iowa Tests of Educational Development™*). Please see **www.lexile.com** for more information.

Linkage - The result of placing two or more tests on the same scale, so that scores can be used interchangeably. Several linkage methods exist. See Equating

Local Norms - Norms (i.e., percentile ranks and stanines) by which test scores are referred to a specific, limited reference population of particular interest to the test user (e.g., norms group is based on state, district, or school data) and are not intended as representative of Resource: Joint Committee on Standards for Educational and Psychological Testing of the AERA, APA, and NCME. (1999). Standards for educational and psychological testing. Washington DC: American Educational Research Association. populations beyond that particular setting. Used to evaluate a student's performance on a test in comparison to other students within the same subpopulation of interest. Similar to NPRs, LPRs provide the percentage of scores for a distribution that falls at or below a given score. Percentile Ranks range in value from 1 to 99, and indicate the status or relative standing of an individual within a specified group (e.g., norms group), by indicating the percent of individuals in that group who obtained lower scores. LPRs can be used, and are used by customers, to evaluate a student's performance on a test in comparison to other students within the same subpopulation of interest. **See also:** Norms, Reference Population, Percentile Rank, and Stanine.

Local Percentile Rank - see Local Norms and Percentile Rank.

Local Stanine - see Local Norms and Stanine.

Longitudinal – Assessment data dealing with the growth or change of an individual or group over time. Refers to a type of reporting style that displays the current test scores and test scores from previous years along with information pertaining to the amount of change or rate of growth (e.g., differences in Developmental Standard Scores, Grade Equivalents, etc.) that is indicated from one year to another. **See also:** Growth, Expected Growth, Developmental Standard Score, Scale Score, and Grade Equivalent.

Mastery Level - The cut score for a criterion-referenced or mastery test. Test takers who score lower than the cut score, or "below the mastery level" are considered not to have mastered the test material, while those scoring at or above the cut score, or "above the mastery level", are considered to have demonstrated mastery of the test material. The method of setting the score designated as representing "mastery" can vary, and is often subjectively determined. **See also:** Cut Score, Criterion-Reference Test, Criterion- Referenced Interpretation, and Mastery Test.

Mastery Test - A criterion-referenced test designed to indicate the extent to which the test taker has mastered a given unit of instruction or a single domain of knowledge or skill. Mastery is considered exemplified by those students attaining a score above a particular cut score (i.e., passing score). **See also:** Criterion-Referenced Test, Criterion-Referenced Interpretation, Cut Score, Mastery Level.

Mean - see Arithmetic Mean.

Measurement - The assignment of numbers to observations or individuals in a systematic manner as a way of representing or quantifying properties or characteristics of each individual. **See also:** Psychometric.

Median - The middle point (score) in a distribution of ranked-ordered scores that divides the group into two equal parts, each part containing 50% of the data. The median corresponds to the 50th percentile. Half of the scores are below the median and half are above it, except when the median itself is one of the obtained scores. For example, the median of the set {4, 7, 8, 9, 10, 10, 12} is 9. **See also:** Percentile.

Mid-Year Norms - Taking into account the date of testing is an important factor in normreferenced interpretations. Students testing during the middle of the school year have not been given the same amount of instruction as those testing in the spring. Because of this, norm tables (specifically National Percentile Ranks) are provided for different times of the Resource: Joint Committee on Standards for Educational and Psychological Testing of the AERA, APA, and NCME. (1999). Standards for educational and psychological testing. Washington DC: American Educational Research Association. year. Mid-Year norms, sometimes referred to as Winter Norms, are typically provided for those students testing from December to February, however this time frame may vary depending on the test that was administered. For more specific information, consult the interpretative guide and technical manual for the test administered. **See also:** Norms, Norm- Referenced Interpretation, and Percentile Rank.

Mode - The score that occurs most frequently in a distribution of scores. For example, the mode for the set {65, 72, 85, 85, 70, 90} is 85. A distribution can have more than one mode. For example, the modes for the set {65, 72, 72, 72, 85, 85, 85, 70, 70, 90} are 72 and 85, since both these scores have a frequency of three, which is the highest frequency for the distribution. **See Figure 8.**

Examples:

Score	Frequency	Score	Frequency
65	1	65	1
70	1	70	2
72	1	72	3*
85	2*	85	3*
90	1	90	1

Figure 8: Mode Examples

Modification - see Test Modification.

Multiple-Choice Item - A type of test item that requires students to select a response (the correct or best answer) from a group of possible choices (response options or alternatives) in determining the answer to the question posed (the stem), which is often related to other information provided (stimulus). Also referred to as a selected-response or forced-choice item. **See also:** Stem, Foil, Prompt (Stimulus).

National Percentile Rank (NPR or PR) - A Percentile Rank indicating the status or relative rank of a student's score compared to a nationally representative sample of examinees. See Percentile Rank and Norms.

National Stanine (NS or Sta9) - A stanine score indicating the status or relative rank of a student's score compared to a nationally representative sample of examinees. See Stanine and Norms.

N-count (N) - A term that refers to the total number of individuals in a particular group (e.g., classroom, building, demographic, etc.) or the total number of data points (scores) in a score distribution entering into a particular statistical calculation. For example, for the set {85, 76, 97, and 66}, N is equal to 4. **See also:** Distribution.

Normal Curve Equivalent (NCE) - Normalized standard scores that range in value from 1 to 99, and have a mean of 50 and a standard deviation of 21.06. In order to interpret these scores it is often necessary to relate them to other scores such as percentile ranks or stanines. NCE's can be thought of as roughly equivalent to stanines to one decimal place. **For example**, an NCE of 48 may be interpreted as a stanine of 4.8. One advantage to NCE's is that they have the properties of an equal interval score, which means they can be averaged, unlike Percentile Ranks. Another advantage to NCE scores is that all test publishers derive these scores in the same way, which leads to some comparability in reporting procedures. **See Figure 9. See also:** Status Scores, Percentile Rank, Standard Score, and Stanine.

NCE-PR relationship		NCE-Stanine relationship	
NCE	PR	NCE	Stanine
99	99	86-99	9
90	97	76-85	8
80	92	66-75	7
70	83	56-65	6
60	68	45-55	5
50	50	35-44	4
40	32	25-34	3
30	17	15-24	2
20	8	1-14	1
10	3		
1	1		

Figure 9: Normal Curve Equivalent Example

Normal Distribution - Also known as the bell-shaped curve be cause of its distinctive appearance in that scores are distributed symmetrically about the middle, such that there are an equal number of scores above as below the mean, with more scores concentrated near the middle than at the extremes. The normal distribution is a theoretical distribution defined by specific mathematical properties that many human traits and psychological characteristics appear to closely approximate (e.g., height, weight, intelligence,etc.). **See Figure 10. See also:** Distribution, Arithmetic Mean, Median, Mode, and Standard Deviation. Some features of the normal distribution are:

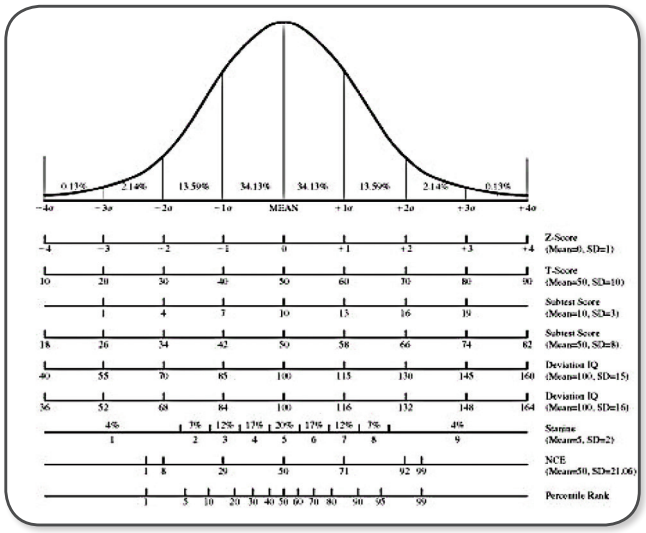


Figure 10: Normal Distribution Example

1.) The mean, median, and mode are identical in value.

2.) The scores are distributed symmetrical about the mean (50.0% above the mean and 50.0% below the mean).

3.) 68.26% of the scores are within 1 standard deviation of the mean (34.13% above the mean and 34.13% below the mean).

4.) 95.44% of the scores are within 2 standard deviations of the mean (47.72% above the mean and 47.72% below the mean).

5.) 99.72% of the scores are within 3 standard deviations of the mean (49.86% above the mean ad 49.86% below the mean).

Norm Group - see Standardization Sample, Norms, Norm-Referenced Interpretation.

Norm-Referenced Interpretation - A score interpretation based on a comparison of a test taker's pe rformance to the performance of other people in a specified reference population (e.g., age groups, grade groups, etc.). Norm-referenced interpretations can be for individuals (i.e., student norms) or for institutions (e.g., school norms) and could involve converting scores to scale scores (or standard scores), percentile ranks, stanines, grade equivalents, etc., depending on use of the test and the information provided by the test publisher. Norm-referenced interpretations allow educators to get an "external" look at the performance of their students in relation to rest of the nation. Contrast to Criterion- Referenced Interpretation. **See also:** Reference Population, Age-Based Norms, Grade-Based Norms, Norm-Referenced Test, Norms, Standard Score, Scale Score, Percentile Rank, Stanine, and Grade-Equivalent.

Norm-Referenced Test (NRT) - Any standardized test or evaluative instrument for which the resulting scores are interpreted or acquire additional meaning in terms of comparisons made to a specified group (i.e., reference group) for which the individual or group belongs (e.g., age or grade). Tests can be considered both norm-referenced and criterion-referenced. It depends only on the use and interpretation of the scores. Contrast to Criterion- Referenced Interpretation and Criterion-Referenced Test. **See also:** Standardized Test, Norm-Referenced Interpretation, and Reference Population.

Norms - Statistics or tabular data that summarize the distribution of test scores for one or more specified groups (e.g., by age or grade groups) that act as a frame of reference for which the performance of subsequent examinees can be gauged. Norms are typically developed using a nationally representative sample (i.e., the standardization sample or norms group) of the group of examinees to be represented by the norms (i.e., the reference population). Norms represent typical performance, not necessarily a desired level of performance (i.e., standards). **See also:** Distribution, Age-Based Norms, Grade-Based Norms, Sample, Standardization Sample, Standardization, Reference Population, Norm- Referenced Test, and Norm-Referenced Interpretation. Contrast to Standards.

Number Attempted (No. Att.) - The number of items that an individual attempts to answer on a test. The same term applies regardless of the type of items (e.g., multiple-choice, constructed-response, etc.) included on the test - the total number of items with non-blank responses.

Objective - A statement of some desired educational outcome - what students are expected to learn at various developmental levels to indicate progress toward meeting specific content standards. A test could measure several different objectives. Examples of objectives are "pattern recognition", "decision making", and "applying scientific concepts to real life problems". **See also:** Content Standard.

Objective Test - A test or assessment that is composed of items for which the correct answers are known in advance of test administration and are unaffected by scoring opinion or judgement. Objective tests typically have "keyed" responses, such that the items can be scored mechanically or by computer. Examples of objective tests are those that contain item formats such as multiple-choice, selected-response, matching, true -false, or other forcedchoice type of item for which there is an indisputable correct answer. Contrast to Subjective Test. **See also:** Item and Multiple-Choice Item.

Off-Grade Testing - An optional test specific to a state that is aligned to that state's learning requirements and/or standards. This test is not required by the state, but is a test developed for the purpose of helping schools measure student's progress in years prior to or after an on-grade testing year. For example, if the state's mandated testing is planned for grades 3, 5, and 8, off-grade tests could be purchased for administration in grades 2, 4, 6, and 7. Compare to On-Grade Testing. **See also:** Standards, Objective, and Criterion-Referenced Test.

On-Grade Testing - A statewide test that is administered to all students in a s pecific grade or grades as required by the state. The tests are aligned to state objective specifications, and the development of materials is done in conjunction with panels of educators and members of the state department of education. Compare to Off-Grade Testing. **See also:** Standards, Objective, and Criterion-Referenced Test.

Open-Ended Response - see Constructed-Response Item.

Outfit - Fit statistics are used to determine how well a particular item or person is 'fitting' with the IRT model being implemented. The outfit statistic is sensitive to the 'extreme' (easy or difficult) persons or items in the data set. This statistic will range from zero to infinity, with the target value being 1.0. **See also:** Infit.

Out-of-Level Testing - The practice of administering a level of a test that is either higher than or lower than the recommended level that is typically based on the student's age or grade placement. The purpose of testing out-of-level is to better match the difficulty of the test to the student's ability or current level of functioning. **See also:** Floor and Ceiling.

Parallel Forms - see Alternate Forms.

Percent Correct (PC) - The percentage of the total number of points that a student received on a test. The percent correct score is obtained by dividing the student's raw score by the total number of points possible and multiplying the result by 100. For multiple-choice tests, this is the same as dividing the student's raw score by the number of questions (i.e., each item is worth one point) and multiplying by 100. Percent Correct scores are typically used in criterion-referenced interpretations and are only helpful if the overall difficulty of the test is known. Often incorrectly interpreted as a Percentile Rank. **See also:** Criterion-Referenced Interpretation.

Percentile - The score or point in a score distribution at or below which a given percentage of scores fall. For example, if 72 percent of the students score at or below a score of 25 on a given test, then the score of 25 would be considered at the 72nd percentile. Contrast to Percentile Rank and Percent Correct scores. **See also:** Distribution, Median, Quartile, and Decile.

Percentile Band - A type of confidence interval, constructed around the examinee's obtained Percentile Rank, indicating with a given probability or confidence, the range of scores in which the examinee's true score may lie. Typically represented on score reports as the shaded area or band, around the obtained Percentile Rank (signified by a diamond or other symbol), extending from one (or more) standard error(s) of measurement below the obtained score to one (or more) standard error(s) of measurement above the obtained score. **See also:** Confidence Interval, Percentile Rank, True Score, and Standard Error of Measurement.

Percentile Rank (PR) - The percentage of scores in a specified distribution that fall at or below the point of a given score. Percentile Ranks range in value from 1 to 99, and indicate the status or relative standing of an individual within a specified group (e.g., norms group), by indicating the percent of individuals in that group who obtained lower scores. For example, if a student earned a 72nd Percentile Rank in Language, this would mean he or she scored better than 72 percent of the students in a particular norm group who were administered that same test of Language. This also implies that the only 28 percent (100 - 72) of the norm group scored the same or higher than this student. Note however, an individual's percentile rank can vary depending on which group is used to determine the ranking. A student is simultaneously a member of many groups: classroom, grade, building, school district, state, and nation. Test developers typically publish different sets of percentile ranks to permit schools to make the most relevant comparisons possible. Contrast to Percentile and Percent Correct scores. **See also:** Status Scores, Distribution, Norms, and Norm-Referenced Interpretation.

Performance Level Descriptor - A set of statements describing a proficiency level, used as a guideline for evaluating student performance by indicating the expectations of what a student at that level can or cannot do. For example, the set of statements describing what it means for a student to be "proficient" in a particular content area. **See also:** Achievement levels/Proficiency levels, Performance Standard, Cut Score, and Criterion-Referenced Interpretation.

Performance Criteria - see Rubric, Performance Standard, Cut Score.

Performance Standard - 1.) A definition of a certain level of performance in some domain in terms of a cut score or a range of scores on the score scale of a test measuring proficiency in that domain. 2.) A statement or description of a set of operational tasks exemplifying a level of performance associated with a more general content standard. The statement may be used to guide judgements about the location of a cut score on a score scale. The term often implies a desired level of performance and is typically established by experts within the field education. **See also:** Cut Score and Content Standard.

Performance Assessment - Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions which specific knowledge or skills are actually applied. In the strictest sense, performance assessments involve some motor or manual response on the part of the examinee and tend to minimize the role of language. However, performance assessments have, within recent years, been used to denote tests that require the examinee to produce a work-sample or written response, and have become synonymous with tests that are composed of mostly constructed-response items. **See also:** Constructed-Response Item.

Pilot Test - A test administered to a representative sample of examinees for the sole purpose of trying out some aspects of the test or test items, such as instructions, time limits, item response formats, or item response options. Typically, a pilot test provides test developers a first empirical look at an item's performance (e.g., difficulty) and is used to make informed decisions on the future development of quality test forms. In some testing programs, the pilot test is also called the field test. Item performance is often evaluated by means, p-values, point biserials, response or distractor analyses, Differential Item Functioning studies, etc. and often times reviewed during a procedure called data review in a collaborative effort between content specialists and psychometricians.

Point Biserial - A statistic that is used as an index of the extent to which an item discriminates between low and high scorers. The point biserial is calculated by determining the correlation between an item and a criterion measure, such as the total test score, and ranges in value from -1.0 to +1.0. Higher values, in general, are more desirable, and those with negative values are considered extremely undesirable. Indicates the extent to which the group's performance on an item is related to their performance on all other items on the test. **See also:** Item Discrimination, Correlation, Correlation Coefficient, and Item.

Portfolio - A systematic collection of education or work products that have been compiled or accumulated over time, according to a specific set of principles.

Power Test – A test that either has no time limit or has a very generous time limit to ensure that each examinee has sufficient time to attempt each item. Power tests are intended to measure the range of an examinee's capacity in a particular area regardless of the speed of the response. These tests tend to be ordered in terms of increasing item difficulty and are generally composed of items that have an adequate level of difficulty, such that even without time limits, only a certain number of items could be answered correctly. Contrast to Speed Test.

Predicted Achievement – An estimate of a student's score on an achievement test (e.g., standard score or scale score) based on statistical estimation methods (prediction equations) using some other test score (e.g., ability score) as the predictor variable. The difference between a student's observed achievement score and predicted achievement score can be examined to determine if the is student performing in accordance to his/her potential, or in terms of expected growth. **See also:** Ability/Achievement Discrepancy, Achievement Test, and Ability Testing.

Predictive Validity - see Validity.

Private School Norms - Norms (specifically Percentile Ranks) that are developed from distributions of student test scores, using a representative sample of students attending private schools only (e.g., preparatory schools, non-catholic religious schools, etc.). Private School Norms are used to indicate the status or relative rank of a student's score compared to other students in the nation who also attend private school. Educational programs for private schools have unique characteristics that affect test performance, and therefore, separate norm tables are provided. See Norms, Percentile Rank, Distribution, and Standardization Sample.

Proctor - Someone who is hired or appointed to supervise students during the administration of an examination. A Proctor's responsibilities may include such activities as the distribution and collection of test materials, monitoring the total testing time, and observing student behavior during the test to ensure that students are not cheating.

Profile - A graphic presentation of the scores for an individual or group, representing the results of several tests (or subtests) that have been expressed in comparable terms (standard scores, percentile ranks, etc.). This type of display is useful for easily identifying relative strengths and weaknesses. As applied to ***CogAT***, the Ability Profile, assigned to students and appears on score reports, is a 9-character code summarizing the level and pattern of the ***CogAT*** scores that is (or can be) depicted graphically. Information from ***CogAT*** Ability Profiles can help teachers adapt instruction, learning materials, and the pace of instruction to the individual needs of their students. For more information on the use and interpretation of ***CogAT*** Ability Profiles see www.cogat.com.

Prompt (Stimulus) - Information presented in a test item (the question, stimulus or instructions) that activates prior knowledge and requires analysis in order for a student to respond. A stimulus or prompt could be a reading passage, map, chart, graph, drawing, photograph, or any combination of these. Most commonly used in item sets, where several items ask questions that refer to the information in the prompt. A terminology generally associated with constructed-response items. **See also:** Item and Constructed-Response Item.

Protocol - A record of events. A test protocol usually consists of the test record and test scores. Typically, a document or form used for recording the results of an individually administered test.

Psychometric - Pertaining to the quantitative measurement of academic, psychological or mental characteristics such as abilities, aptitudes, knowledge, skills, and traits. **See also:** Measurement. and can be examined at multiple levels, including item, subtest, domain, overall test, etc. Examination of the psychometric characteristics at each of these levels provides useful insight into the measurement properties of assessments.

P-value (Item Difficulty Index) - An index, designated as p (or p-value), that indicates an item's difficulty, calculated as the proportion of some specified group, such as grade, who answer a test item correctly. P-values range in value from 0.0 to 1.0, with lower values corresponding to more difficult items and higher values corresponding to easier items. Typically used in reference to multiple-choice items or other item types in which the item is only worth 1 point. For constructed-response items or item types in which the item is worth more than 1 point, a p-value can be estimated by dividing the resulting item mean by the maximum number of points possible for the item. **See also:** Item Difficulty, Item, Multiple- Choice Item, and Constructed-Response Item.

Quartermonth Norms - A quartermonth is a unit of time equal to one quarter of a month. Used specifically in reference to interpolated norms tables. Typically norms are developed from data gathered at one or two points during the school year (a testing window in the fall and/or spring) and the normative information for the time interval between these points is approximated by interpolation. Assuming a traditional 10-month school year (September through June), there are at most forty quartermonth norm tables available. Thus, providing a better estimate of normative information throughout the school year (based on the amount of instruction) for those schools that desire the flexibility of testing at any time during the year (outside the empirical norm window). **See also:** Norms, Interpolated Norms, Interpolation, Empirical Norms Dates.

Quartile - One of three points (defined as low, middle, or upper) which divide the scores in a distribution into four equal groups, each containing 25% of the data. Quartiles are special cases of percentiles -the lower, middle, and upper quartiles correspond to the 25th, 50th (median), and 75th percentiles. **See also:** Median, Percentile, and Interquartile Range.

Random Error - An unsystematic measurement error. A quantity (often observed indirectly) that appears to have no relationship to any other variable or is due to random causes. Contrast to Systematic Error.

Random Sample - see Sample.

Range - The range of a distribution of scores is defined as the difference between the two extremes (maximum score minus minimum score), and is a rough indication of the spread or variability of the scores. The range of the set {65, 72, 85, 85, 85, 70, 90} is calculated as (90 - 65), which is equal to 25. **See also:** Variability.

Rasch – A type of IRT model used to analyze data from assessments. The Rasch model is a special case of a 1-parameter model within the framework of Item Response Theory.

Rating Scale – A scale whose points are defined by predetermined criteria (verbal, numeric, or symbolic descriptors), and with which judgements concerning the strength of a particular trait, along a continuum, are indicated. Rating scales are one method for collecting and quantifying certain subjective judgements, such as gathering self-rating information on inventory or psychological assessments (e.g., personality assessments), or systematizing the gathering of expert judgements across many raters for a given purpose (e.g., rubrics for scoring performance assessments). **See also:** Inventory, Performance Assessment, and Rubric.

Raw Score (RS) - The first unadjusted score obtained in scoring a test. A Raw Score is usually determined by tallying the number of questions answered correctly or by the sum or combination of the item scores (i.e., points). However, a raw score could also refer to any number directly obtained by the test administration (e.g., raw score derived by formulascoring, amount of time required to perform a task, the number of errors, etc.). In individually administered tests, raw scores could also include points credited for items below the basal. Raw Scores typically have little meaning by themselves. Interpretation of Raw Scores requires additional information such as the number of items on the test, the difficulty of the test items, norm-referenced information (e.g., Percentile Ranks, Grade Equivalents, Stanines, etc.), and/or criterion-referenced information (e.g., cut -scores). Often times in standardized tests, raw scores are converted to standard scores or scale scores for reporting purposes (see Scale Scores). **See also:** Item, Item Difficulty, Norm-Referenced Interpretation, Criterion-Reference Interpretation, Basal, Percentile Rank, Grade Equivalent, Stanine.

Readability – In a test development context, a term that refers to the difficulty level of a reading passage or other text related to a test or test item. There are several different formulas for calculating the readability of a text including Flesch, Kincaid-Flesch, Dale- Chall, Gunning-Fog, Raygor, and Fog. Information such as the total number of words in the reading selection, average sentence length, number of sentences, length of words (e.g., number of syllables), etc., is examined through the use of computer algorithms; and a number, referred to as the readability index, is assigned indicating the grade -level appropriateness of the text. Measures of readability are useful in the selection of grade - appropriate reading passages and test directions.

Readiness Test - A type of prognostic test that is us ed to predict a student's future success in undertaking or engaging in a new learning activity. The test measures the student's physical, mental, or emotional preparedness or maturity to handle formal instruction. For example, a test of reading readiness determines whether or not student has acquired the appropriate skills or attained the proper developmental stage to begin to learn to read. **See also:** Aptitude Test.

Reference Population - The population of test takers represented by the test norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of examinee age, grade, clinical status, or other characteristics at the time of testing (e.g., school type - Catholic or Private norms). **See also:** Norms and Sample.

Regression to the Mean - The tendency of extreme scores on one test to be less extreme on a second related test. If a person is administered two tests that are related, but not perfectly correlated (e.g., intelligence and achievement tests), and the person scores above (or below) the average score on one test (e.g., intelligence test), on the average the person will most likely score above (or below) the average on the second test, but not as far above (or below) as on the first test. The second score "moves" or "regresses" to the mean or average score. The amount of regression is dependent on the degree of correlation between the two tests. The greatest regression effect occurs if two tests have a zero correlation (McGrew, 1994). **See also:** Correlation, and Average.

Reliability - The degree to which test scores for a group of examinees are consistent over repeated administrations of the same test, and therefore considered dependable and repeatable for an individual examinee. A test that produces highly consistent, stable results (i.e., relative free from random error) is said to be highly reliable. The reliability of a test is typically expressed as a reliability coefficient or by the standard error of measurement derived by that coefficient. **See also:** Reliability Coefficient, Standard Error of Measurement, and Random Error.

Reliability Coefficient - A statistic that reflects the degree to which scores are free of measurement error. This statistic is often expressed as a correlation coefficient (e.g., correlation between two forms of a test - Alternate-Forms reliability, or between two administrations of the same test - Test-Retest reliability), or resembles a correlation coefficient (e.g., calculation of the test's internal consistency using K-R 20 or coefficient alpha). With each of these reliability coefficients, the higher the value of the index (closer to 1.0), the greater the reliability of the test. **See also:** Reliability, Internal Consistency, Error of Measurement, Alternate Forms, Test-Retest Reliability, and K-R 20.

Representative Sample - see Sample.

Rubric - An established set of criteria, including rules, principles and illustrations used to determine the caliber of a student's performance on an assessment task or constructedresponse item. Scoring rubrics often vary in terms of the degree of judgement entailed, in the number of distinct score levels defined, and in the latitude given to scorers for assigning intermediate or fractional score values. **See also:** Constructed-Response Item.

Sample - A selection of a specified number of entities, sampling units (e.g., examinees, items, etc.) from a larger specified set of possible entities, the population. A random sample is a selection of entities from the population according to a random process in which the each entity has a equal chance of being selected and the selection of one entity from the population is in no way related (or dependent) to the selection of other entities. A stratified random sample is a set of random samples, each of a specified size, from several different sets, in which the sampling is viewed as or is representative of the strata (levels of defined categories) of the population. Representative samples are samples that correspond to or match the population from which it was drawn with respect to characteristics considered important for the purposes under investigation. For example, nationally representative Resource: Joint Committee on Standards for Educational and Psychological Testing of the AERA, APA, and NCME. (1999). Standards for educational and psychological testing. Washington DC: American Educational Research Association. norm samples are typically "stratified" according to information from the national census (e.g., proportion of gender/ethnicity, school type, regional information, and socioeconomic status). **See also:** Standardization Sample.

Scale Score (Scaled Score, SS) - A type of derived score, which is a transformation of the raw score, developed through a process called scaling. Scale Scores provide a continuous score scale (developmental scale) across different levels and forms of a test that permits the direct comparison of different groups of examinees - regardless of the time of year tested and the level/form administered (i.e., useful for longitudinal comparisons). Scale Scores, unlike Percentile Ranks and Grade Equivalents, are equal-interval, a property that allows these scores to be added, subtracted, and averaged. The term scale score and standard score are often used interchangeably, even though these scores may be derived at by different methods, their purpose and use can be similar. **See also:** Derived Score, Raw Score, Scaling, Standard Score, Universal Scale Score. Contrast to Percentile Rank and Grade Equivalent.

Scaling - The process of creating a scale or a scale score. Scaling may enhance test score interpretation by placing scores from different tests or test forms onto a common scale (unit of measurement) or by producing scale scores designed to support criterion-referenced or norm-referenced score interpretations. Methods for scaling vary and interpretations are typically specific to the testing program and don't generalize across the testing program. One type of scaling that helps improve interpretations across grades or levels is vertical scaling. Vertical scaling is a process that results in scales that are interpretable across grades and levels and can be useful for examining growth or cohort comparisons (see vertical scaling). **See also:** Scale Score, Criterion-Referenced Interpretation, and Norm-Referenced Interpretation.

School Norms - Norms (specifically National Percentile Ranks) that are developed from distributions of average student scores, using a nationally representative sample of students. The average scores are computed for each grade within a school building. School norms are used to indicate the status or relative rank of the school (based on average student performance per grade) compared to other schools in the nation. See Norms, Percentile Rank, Status Scores, Distribution, and Standardization Sample.

Screening Test - A test that is used to quickly and efficiently make broad categorizations of examinees or to identify individuals who deviate in a specified area, as a first step in selection decisions (e.g., readiness for acade mic work) or diagnostic processes (e.g., incidence of maladjustment).

Selected-Response Item – A type of item format, usually called a "multiple-choice" item, which requires the test taker to select a response from a group of possible choices, one of which is the correct answer to the question posed. **See also:** Multiple-Choice Item.

Short-Answer Item - An item test format or question that requires only a few words, phrase, or a number as an answer. **See also:** Constructed-Response Item.

Speededness – A test characteristic, dictated by the test's time limits, that results in a test taker's score being dependent on the rate at which work is performed as well as the correctness of responses. One indication of speededness is the percent of test takers completing the test. **See also:** Speed Test and Completion Rates.

Speed Test – A test in which performance is primarily measured by the time to perform a specified task or by the number of tasks performed in an allotted amount of time. A speed test also refers to a test scored for accuracy, while the test taker works under time pressure. Typing tests and tests of reading speed (e.g., number of words per minute) are two examples of speed tests. In an educational testing context, the item difficulties of a speed test are generally such that given no specified time limit, all test takers should be able to complete all test items correctly. Contrast to Power Test. **See also:** Speededness.

Spiraling (Spiraled): A packaging process used when multiple forms of a test exist and it is desired that each be tested in all classrooms participating in the testing process. This process allows for the random distribution of test booklets to students. For example, if a package has 3 copies each of the test booklets (test forms) labeled A, B, C, & D, the order of the test booklets in the package would be: A, B, C, D, A, B, C, D, A, B, C, & D.

Split-half Reliability Coefficient – An internal consistency (reliability) coefficient obtained by using half the items on the test to yield one score, and the other half of the items on the test to yield a second independent score. The correlation between the two scores on these two half-tests, adjusted via the Spearman-Brown formula, provides an estimate of the alternate-form reliability of the total test. The definition of "test halves" varies and could be defined in such ways as "1st half of the test vs. 2nd half of the test" or "odd-numbered items vs. even-numbered items". This measure of reliability is typically inappropriate for speeded tests. **See also:** Reliability Coefficient, Reliability, Internal Consistency, Speededness, Alternate Forms, and Correlation.

Spring Norms - Taking into account the date of testing is an important factor in normreferenced interpretations. Students testing in the spring of a given grade level have been given more instruction than tested in the fall. Because of this, norm tables (specifically National Percentile Ranks) are provided for different times of the year. Spring norms are typically provided for those students testing from March to June, however this time frame may vary depending on the test that was administered. For more specific information, consult the interpretative guide and technical manual for the test administered. **See also:** Norms, Norm-Referenced Interpretation, and Percentile Rank.

Standard Age Scores (SAS) - Normalized standard scores, having a mean of 100 and a standard deviation of 16, provided for each battery and composite on the Cognitive Abilities Test (CogAT). These scores are developed for the purpose of comparing the rate and level of cognitive development of an individual to other students in the same age group. The Standard Age Score be converted to other derived scores such as Age Percentile Rank (APR) and Age Stanine (AS) through the use of a set of conversion tables. **See also:** Standard Score, Norms, Age-Based Norms, Derived Score, Percentile Rank, Stanine, and Universal Scale Score.

Standard Deviation (SD) - A statistic that measures the degree of spread or dispersion of a set of scores. The value of this statistic is always greater than or equal to zero. If all of the scores in a distribution are identical, the standard deviation is equal to zero. The further the scores are away from each other in value, the greater the standard deviation. This statistic is calculated from using the information about the deviations (distances) between each score and the distribution's mean. It is equivalent to the square root of the variance statistic. The standard deviation is often the preferred me thod of examining a distribution's variability since the standard deviation is expressed in the same units as the data. **See also:** Variability, Variance and Arithmetic Mean.

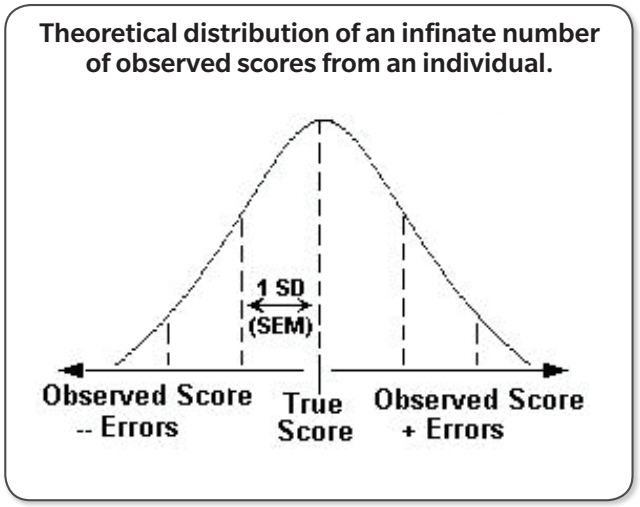


Figure 11: SEM Example

Standard Error of Measurement (SEM) - A statistic used to indicate the amount of expected error in a score. Theoretically, this can be thought of as the standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of test) under identical conditions. Alternatively the SEM can be thought of as the standard deviation of the errors of measurement (true score - observed score) from an infinite number of repeated test administrations. Because such data cannot be generally collected (i.e., the true score is not known and infinite test administrations are not possible), the standard error of measurement is estimated from group data using information about the reliability and the standard deviation for a set of test scores. The smaller the calculated standard error of measurement, the greater the accuracy of the score. The SEM is often used in the construction of a confidence interval around an individual's observed test score. **See Figure 11. See also:** Confidence Interval, True Score, Error of Measurement, Reliability, Standard Deviation, Alternate Forms.

Standardization - 1.) In test administration, maintaining a constant testing environment and conducting the test according to detailed rules and specifications, so that testing conditions are the same for all test takers (e.g., testing materials, directions for administration, time limits, etc.). 2.) In test development, establishing scoring norms based on the test performance of a nationally representative sample of individuals (i.e., standardization sample) for which the test is intended for future use. These individuals are administered the test using the same set of instructions, testing materials, time limits, etc. as is intended for the operational use of the test. **See also:** Standardization Sample, Standardized Test, Sample, Norms, Norm-Referenced Test, and Norm-Reference Interpretation.

Standardization Sample - A sample of students from the reference population whose resulting test scores are used in the development of norms. **See also:** Sample and Reference Population, Standardization, Norms, Norm-Referenced Test, and Norm-Referenced Interpretation.

Standardized Test - A test designed to be administered, scored, and interpreted according to a prescribed set of rules or instructions. The instructions for administration require that the testing conditions are the same for all examinees so that the results can be interpreted through comparisons to a specific reference group (norm group) who were administered the test under the same conditions. **See also:** Standardization, Standardiza tion Sample, Reference Population, Norms, Norm-Referenced Test, and Norm-Referenced Interpretation.

Standards - Statements of what students are expected to learn, or should be able to do, as the intended results of their experience with the curriculum for a given subject area. **See also:** Content Standards.

Standards-Based Assessment - Assessments intended to represent systematically described content and performance standards. **See also:** Content Standard, Performance Standard, Standards.

Standard Score (SS) - A type of derived score, which is a transformation of the raw score, and whose score distribution in a specified population has convenient, known values for the mean and standard deviation. Often this term is used to specifically denote z-scores (mean =0.0 and standard deviation =1.0), and any linear transformation of z-scores. However, Standard Scores can also be developed to provide a continuous score scale (developmental scale) across different levels and forms of a test. Standard Scores permit the direct comparison of examinees by placement of the scores on a common scale and, for this reason, are useful for longitudinal comparisons. The term scale score and standard score are often used interchangeably, even though these scores may be derived at by different methods, their purpose and use can be similar. **See also:** Derived Score, Raw Score, Arithmetic Mean, Standard Deviation, Z-Score, Scale Score, Development Standard Score, and Standard Age Score.

Standard Setting - The procedure used in the determination of the cut scores for a given assessment that is used to measure students' progress towards certain performance standards. Standard Setting methods vary (e.g., item-centered, such as Bookmark or holistic/person-centered, such as Body of Work), but most typically employ the use of a panel of educators and expert judgements, rounds of judgments, and empirical data to operationalize the level of achievement students must demonstrate in order to be categorized within each performance level. **See also:** Cut Score, Performance Standard, Standards, Standards -Based Assessment, Criterion-Referenced Interpretation, and Achievement Levels/Proficiency Levels.

Stanine - The name stanine is simply a derivation of the term "standard-nine" scale. Stanines are normalized standard scores, ranging in value from 1-9, whose distribution has a mean of 5 and a standard deviation of 2. Stanines 2 through 8, are equal to a ½ standard deviation unit in width, with the middle stanine of 5 defined as the range of scores ¼ of a standard deviation below to ¼ of a standard deviation above the mean. Stanines can, more easily, be thought of as coarse groupings of percentile ranks (see below), and like percentile ranks indicate the status or relative rank of a score within a particular group. Due to their coarseness, stanines are less precise indicators than percentile ranks, and at times be misleading (e.g., similar PR's can be grouped into different stanines [e.g., PR=23 and PR=24] and dissimilar PR's can be grouped into the same stanine [e.g., PR=24 and PR=40]). However, some find that using stanines tends to minimize the apparent importance of minor score fluctuations, and are often helpful in the determination of areas of strength and weakness. **See Figure 12. See also:** Standard Score, Status Scores, Percentile Rank, Arithmetic Mean, and Standard Deviation.

Percentile Rank Range	Stanine	Percent of Examinees	Descriptor
96-99	9	4%	high
89-95	8	7%	well above average
77-88	7	12%	above average
60-76	6	17%	somewhat above average
41-59	5	20%	about average
24-40	4	17%	somewhat below average
12-23	3	12%	below average
5-11	2	7%	well below average
1-4	1	4%	low

Figure 12: Stanine Example

Status Scores - A type of score that indicates how a student's test performance (i.e., score) compares with that of others in a given reference group - the class, district, state, nation, etc. Status or relative rank within a single reference group is usually expressed in terms of percentile ranks, stanines, or normal curve equivalents. **See also:** Reference Population, Percentile Rank, Stanine, and Normal Curve Equivalent.

Stem - The item question or problem statement.

Stimulus - see Prompt.

Stratified Sample - see Sample.

Student Norms (**Pupil Norms**) - Norms (specifically National Percentile Ranks) that are developed from distributions of student test scores using a nationally representative sample of students. Student norms are used to indicate the status or relative rank of a student's score compared to other students in the nation, with similar characteristics, such as age or grade. Consequently, local student norms can also be computed using the cohort of students that took a particular administration (see Local Norms). See Norms, Percentile Rank, Status Scores, Distribution, and Standardization Sample.

Student Growth Percentiles (SGPs): Student Growth Percentiles express growth normatively. They describe the degree to which a student has grown relative to peers with similar past test scores. SGPs are a growth measure that can be used for evaluating student progress and achievement to determine whether or not they are eligible for or would benefit from supplemental educational services and for examining strengths and weaknesses of academic programs to develop instructional action plans to ensure academic excellence.

Subject Area - A body of content derived from related disciplines and organized for curriculum. Some examples of subject areas include English language arts, mathematics, social studies, and science.

Subjective Test - A test or assessment for which some subjectivity on the part of the examiner or rater (as in a panel of score judges) is inherent in the determination of an examinee's scored data. Projective devices, free response examinations, holistic evaluations, observational data, all require some opinion or judgement on the part of the rater in the assignment of scores, even if "objective" scoring guidelines are included as part of the process. Contrast to Objective Test. **See also:** Holistic Scoring and Rubric.

Summative Assessment – The use of tools and techniques to measure student understanding at the completion of teacher instruction. Summative assessment, in contrast to formative assessment, is primarily used by teachers for evaluation purposes (e.g. grades) and not for identification for remediation.

Systematic Error - A consistent score component or measurement error (often observed indirectly), that is not related to the test performance. Also known as a bias. Contrast to Random Error.

Technical Advisory Committee - A group of individuals (e.g., professionals in the field of education, testing, etc.) that are either appointed or selected to make recommendations for and to guide the technical development of a given testing program (e.g., state testing program for a criterion-referenced test).

Technology Enhanced Item (TEI) - An item format, designed in a specific way that requires examinees to utilize computer functionality (drag and drop, hot spots, creating graphs and plots, categorize or classify, fill in the blanks, etc.) in a way that cannot be achieved with a multiple choice or constructed response item. TEIs take advantage of the interactive format to create specialized interactions for collecting response data and to assess relevant knowledge and skills as required in the assessed ability or achievement standards.Test Analysis Sheet knowledge and skills as required in the assessed ability or achievement standardsion purposes (e.g. grades) and not for identificatioem identifier, content area, grade level, standard alignment information, prompt/scenario association, and additional attributes. In the case of tests containing items which have been field tested, additional data such as p-value, item-total correlation, and IRT statistics are included.

Test Battery - see Battery.

Test Coordinator's Manual (TCM) - A publication prepared by test developers and publishers to provide information for test coordinators on the administration of tests.

Test Item - see Item. Computer Enhanced Item - An item format, designed in a specific way that requires examinees to utilize computer functionality (drag and drop, hot spots, etc...) in a way that cannot be achieved with a multiple choice or constructed response item.

Test Modification - A change made to the content, format, and/or administration procedure of a test in order to accommodate test takers who are unable to take the original test under standard test conditions. A testing modification, as distinguished from an accommodation, involves changing the assessment itself so that the tasks or questions presented are different from those used in the regular assessment. This category of change include s modifying test items or the way they are presented so that what is being measured has actually changed. For example, if deaf students were to be signed the listening test, this presentation would change what is being measured. There are different skills required in understanding sign language, compared to comprehending oral discourse. To cite another example of a testing modification, a brailled version of a test modifies the questions just like a translation to another language might. Some test items can only be presented in braille in ways that make responding to be a much more difficult task than with the original test form. Other test items might even have to be eliminated, thereby eliminating part of the test from scoring, because brailling is not possible. When modifications in testing are used in response to a student's disability, special norms may be appropriate, and the interpretation of test scores may need to be qualified. The Individuals with Disabilities Education Act (IDEA) is very clear in cautioning about procedural changes that change the "construct" being measured by an assessment. Often when modifications are used, the meaning of the test scores can change substantially. Notwithstanding the fact that the intent in creating a braille edition is to preserve the original test content and processes as closely as possible, from a measurement perspective, what might become distorted are judgments made about a student's status among peers, based on percentile ranks, or about a student's developmental level, based on grade equivalents. Fortunately, there are useful score interpretations separate from making interindividual comparisons. Percentile ranks can be used to identify relative strengths and weaknesses, and growth can be estimated as long as the same modifications had been used the previous year. The intended use of test scores should always be a prime consideration in reporting decisions based on an administration of a braille edition as well as with any other nonstandard test administration. To return to a previously cited general example, assistance in signing or cueing test items represents a modification. If the assistance involves translating words to another language or providing meanings of words used in the test, these are also modifications. Such assistance may be needed by some ELL students or those whose language development has been slowed by certain disabilities. On a case-by-case basis, the amount of assistance supplied is the key to evaluating the extent to which the modification might interfere with score interpretations and the applicability of the published norms. If an ELL student is given help with the meaning of a few words on a mathematics test, the impact would be much different than if the student is given help with a few words on every item on that test. **See also:** Accommodation, Norms, Norm-Reference Interpretation, Percentile Rank, Grade Equivalent, and Growth.

Test-Retest Reliability - A reliability coefficient obtained by administering the same test a second time to the same group of examinees after a short time interval and correlating the two sets of scores. **See also:** Reliability, Correlation, and Correlation Coefficient.

Test Specifications - A detailed description for a test, often called a blueprint, that specifies the number or proportion of items that assess each content and process/skill area; the format of items, responses, scoring rubrics and procedures; and the desired psychometric properties of the items and test such as the distribution of item difficulty and discrimination indices. **See also:** Rubric, Item Difficulty, and Item Discrimination.

True Score - In Classical Test Theory, the average of the scores that would be earned by an individual on an unlimited number of perfectly parallel forms of the same test. In Item Response Theory, true score refers to the error-free value of test taker's proficiency usually symbolized by theta. Theta is a theoretical construct referring to a test taker's genuine ability, a measure of which could be obtained, in theory, if measurement of such ability through testing could be obtained completely without error - an error-free score. **See also:** Classical Test Theory, Item Response Theory, Error of Measurement, and Alternate Forms.

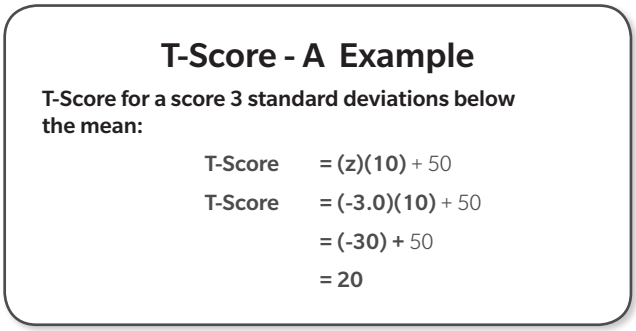


Figure 13: T-Score Example

T-Score - A normalized standard score, having a mean of 50 and a standard deviation of 10. T-Scores are a direct transformation of z-scores and range (roughly) from 20 to 80 (corresponding to approximately 3 standard deviations above and below the mean). **See Figure 13. See also:** Standard Score, Arithmetic Mean, Standard Deviation, and Z-score.

Universal Scale Score (USS) - The Universal Scale Score provides a continuous growth scale of cognitive development from kindergarten through grade 12 for the Cognitive Abilities Test (CogAT). The USS is the standard score that is used for entry to the age and grade norms tables to obtain such derived scores as the Standard Age Score (SAS) and the Grade Percentile Rank (GPR) for each battery and composite score. These scores can then be converted to other derived scores such as Age Percentile Rank (APR), Age Stanine (AS), and Grade Stanine(GS) by the use of another set of conversion tables. **See also:** Scale Score, Age-Based Norms, Grade-Based Norms, Derived Score, Standard Age Score, Percentile Rank, and Stanine.

Validity - The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by the purposed uses of a test. Or more commonly defined, the extent to which the test measures what it is intended to measure (i.e., the accuracy of the test). There are various ways of assessing validity, depending on the type of test and its intended use.

Content Validity **-** The extent to which a test represents a balanced and adequate sampling of the content domain in terms of the knowledge, skills, objectives, etc. (logical/sampling validity) or if the test appears on the surface (rather subjectively) to measure what it is intended to measure (face validity) (e.g., using arithmetic problems to measure quantitative skills).

Criterion-Related Validity **-** The extent to which a test is a measure of a particular criterion measure, either the accuracy of the test to predict performance on some future criterion measure (predictive validity) or is in agreement with some current criterion measure (concurrent validity). Typically, the evidence for this type of validity is determined by calculating a statistic called the validity coefficient (i.e., correlation between the test and the criterion measure (behavior the test is intended to measure)). **See also:** Validity Coefficient, Correlation, and Correlation Coefficient.

Construct-Related Validity **-** The extent to which a test measures the theoretical construct or trait that it is intended to measure (e.g., abstract psychological trait). Often this type of validity is demonstrated by examining the interrelationship of the scores (i.e., correlation) on one test with scores on other tests that are theorized to measure either the same trait or an unrelated trait, and determining if the results are in the expected direction (e.g., high correlation with same trait measures and low correlation with unrelated trait measures).

Consequential Validity **–** This modern approach to validity contends that a major determinant of the validity of an assessme nt is the consequence that the test administration and any subsequent inferences (i.e., score interpretations) have upon the student, instruction, and the curriculum. This approach places heavy emphasis on accruing evidence on both the positive and adverse consequences of testing, whether these consequences were intended or not. Studies of consequential validity focus on such elements as the drop-out rate, curricular changes, teacher retention, improvements in student learning, improvements in the testing program, validation of cut scores, and bias in scoring and interpretation, etc.

Validity Coefficient - Refers to the correlation coefficient that is used as an indication of a test's criterion-related validity. Calculated by determining the correlation be tween the test (predictor) and a criterion score (indicating the behavior the test is intended to measure). For example, a validity coefficient for a college entrance exam could be calculated as the correlation between the student's scores on the exam and their grade point average after four years in college. **See also:** Correlation, Correlation Coefficient, and Validity.

Criterion-Related Validity. Variability - Refers to the degree of spread or dispersion of a set of scores. The closer all the scores in a distribution are to equaling the same value, the less variability the distribution exhibits. The further the scores are away from each other in value, the more the variability. Two statistics most commonly used as a measure of a distribution's variability are the Standard Deviation and Variance. **See also:** Range and Interquartile Range.

Variance - A statistic that measures the degree of spread or dispersion of a set of scores. The value of this statistic is always greater than or equal to zero. If all of the scores in a distribution are identical, the variance is equal to zero. The further the scores are away from each other in value, the greater the variance. The square root of the variance is called the standard deviation, which is often used more frequently than the variance, since the standard deviation is expressed in the same units as the data (as opposed to the squared units of the variance statistic). **See also:** Variability and Standard Deviation.

Vertical Scaling - Vertical scaling can be conceptualized as a measurement process of placing achievement test scores within the same subject but at different grade levels onto a common scale . . . a vertical scale. The implementation of a vertical scale provides educators with a direct measure of students' achievement growth across years. The horizontal equating places scores from different forms of the same test on a common metric from year to year. The statistical methods that are typically used to establish a vertical scale are similar to those applied to equating. However, vertical scaling has two substantive differences. Equating is implemented to the test forms of the same test that are assumed to be assessing the same content at the same level difficulty. In contrast, vertical scaling deals with test forms designed for students at different grades that necessarily differ in content and difficulty, reflecting the difference in grade-to-grade curriculum and student ability. Sometimes the curriculum and content variations between nonadjacent grades may be profound, which poses a problem for the process of vertical scaling that generally does not occur in equating. In order to create a vertical scale, a proper data collection design and scaling method should be selected. One method, the common-item design, uses vertical-linking items appearing in adjacent grades to link scores from these adjacent levels. Through this method, scores from all grade levels are placed onto a base level using a chain-linking process or are simultaneously estimated for all grades.

W-Difference Score - The Relative Performance Inde x (RPI), Standard Score, and Percentile Rank, in any of the tests Dr. Richard Woodcock authors, are based on test or cluster W-difference scores. The W-difference score is the difference between the examinee's test (or Cluster W-Score) and the average test (or cluster W-score) for the reference group in the norm sample (same age or grade) with which the comparison is being made. **See also:** Standard Score, Percentile Rank, W-Score, Average, Reference Population, and Standardization Sample.

W-Score - A special transformation of the Rasch ability scale (Rasch, 1960; Wright & Stone, 1979). The Rasch model is a special case of a 1-parameter model within the framework of Item Response Theory. W-scores are used in tests in which Dr. Richard Woodcock is an author. The W-scale has mathematical qualities that make it well suited as an intermediate step in the interpretation of test performance. It is an equal-interval measurement that is centered at a value of 500, which is set to approximate the average performance of 10-year old individuals. Any cluster score is the arithmetic mean of the W-Scores of the test in that cluster. **See also:** Item Response Theory and Arithmetic Mean.

Weighting - A process of assigning a numeric value called a "weight" to a score, or other variable of interest, to indicate its relative importance to a score distribution or contribution to the calculation of another score or variable. Often the calculation of a composite score involves some differential weighting of the subtest scores that contribute to the composite score, depending on the desired relative impact of these scores on the composite. Another example, is weighting a score distribution to have desired characteristics in terms of the contribution of each group (e.g., defined by regions of the country, school size, gender/ethnic categories, or other pertinent demographic data) to the total sample, with respect to specific target information (e.g., national census data). **See also:** Composite Score.

$$Z = \frac{X - \bar{X}}{SD} = \frac{(25 - 15)}{5} = 2$$

Figure 14: Z-score Example

Z-score - A type of standard score such that the distribution of the scores for a specified population have a mean of 0.0 and a standard deviation of 1.0. The z-score indicates the amount a student's score (X) deviates from the mean in relation to the standard deviation (SD) of the group. For example, if a student's score is 25, and the group's mean and standard deviation are 15 and 5, respectively, then the student's z-score would be 2.0, indicating that the student scored 2 standard deviations above the groupmean. **See Figure 14. See also:** Standard Score, Arithmetic Mean, Standard Deviation.

NOTES

Houghton Mifflin Harcourt
One Pierce Place, Suite 900W
Itasca, IL 60143

Connect with us:

hmhco.com • 800.323.9540

Houghton Mifflin Harcourt

hmhco.com/HMHAssessments