Academic Assessment Terminology

February 2020

Acknowledgement

City Year would like to thank the Center for Assessment for its guidance of this effort to improve how City Year staff understand, interpret, and use assessment results to support students, measure impact, and enable continuous improvement of services provided.

Assessment Types and Uses

"When the cook tastes the soup, that's formative assessment. When the customer tastes the soup, that's summative assessment." – Paul Black

Formative - In-process evaluations of student learning that are typically administered multiple times during a unit, course, or academic program. An exit ticket is one example. The general purpose of formative assessment is to give educators in-process feedback about what students are learning or not learning so that instructional approaches, teaching materials, and academic support can be modified accordingly. Formative assessments are usually not scored or graded and should not be aggregated or compared. Similarly, these results should not be used for evaluation, grades or accountability purposes because of the risk of corrupting the formative assessment process. Effective formative assessment occurs frequently; covers small units of instruction; and is tailored to the student, unit, and classroom. Also known as a **Progress Monitoring** tool.

Interim - An assessment to measure academic status at different points during the school year, and to calculate academic growth between those points to inform instruction during the year. Often designed to measure academic status irrespective of the grade level at which a student is performing. Also known as a **Benchmark** Assessment and sometimes referred to by educators as "**Common Formative Assessments**". Standardized interim assessments may be used to predict a student's likelihood of success on a large-scale summative assessment, evaluate a particular educational program or pedagogy, identify potential gaps in a student's learning after a limited period of instruction has been completed, or measure student learning over time. Some people may label certain teacher assessments "interim", yet usually we prefer to call those teacher-developed assessments or classroom assessments to clearly distinguish them.

Summative - Assessments designed to measure the culmination of a student's learning and typically administered at the end of the school year for accountability purposes. Standardized state-wide assessments fit in this category and are often called **High-Stakes Tests**. Typically, the stakes are high for the school, not the student, although in some states passing this assessment is tied to student graduation. Classroom summative assessments are also in this category. Common examples in middle- and high-school classrooms are "mid-terms," "final projects," and "final exams."

Common Assessments at City Year

MAP - A computer-adaptive assessment from NWEA. An interim assessment typically administered fall, winter, and spring. MAP is aligned to a state's content standards that are published on each state education agency website. MAP aims to measure a student's academic growth, achievement status to state standards, and project proficiency to state standards. MAP assessments are based on Item Response Theory (IRT), a theory of measurement. The numerical value (RIT) assigned to a student represents the most difficult question that he or she is capable of answering correctly about 50% of the time.

STAR – A computer-adaptive assessment from Renaissance Learning, an assessment somewhere between a progress monitoring tool and high-stakes assessment, built to be used for universal screening, progress monitoring and goal setting. Each assessment provides estimates of students' skills and comparisons of students' abilities to national norms. Each is intended to aid with developing curriculum and instruction by providing feedback about student, classroom, and grade level progress. The software reports grade equivalents, percentile ranks, and normal curve equivalents. (Careful, don't confuse this with STAAR, the State of Texas Assessments of Academic Readiness or California's STAR, Standardized Testing and Reporting)

iReady Diagnostic – A computer-adaptive interim assessment from Curriculum Associates for use in kindergarten through high school. The assessment is built on the College- and Career-Ready Standards and tied to online instructional modules. Test scores are intended to pin-point student ability, level of mastery, and progress monitoring over time at subject and subskill levels. iReady also flags students who rushed through the test and may need a retest, and provides growth targets that include information about what typical and stretch growth looks like for each student.

HMH Reading Inventory (RI) - A computer-adaptive reading assessment from Houghton Mifflin Harcourt with foundational and comprehensive subtests for use in kindergarten through high school. The comprehension assessment is linked to the Lexile® Framework for Reading which allows RI to borrow growth information from the Lexile® test scale. The foundational subtest was originally designed for reading level placement decisions and the test developer recommends using additional measures to effectively monitor progress. Through the Lexile® link, RI also provides information that helps match readers to appropriate texts based on their test results

HMH Math Inventory (MI) - A computer-adaptive mathematics assessment from Houghton Mifflin Harcourt for use in screening and growth monitoring in kindergarten through Algebra II. The MI assessment is linked to the Quantile® Framework for Mathematics which allows MI to borrow growth information from the Quantile® test scale.

Fountas & Pinnell Benchmark Assessment System (BAS) – An assessment administered in-person that determines a student's reading level. The reading level is defined based on a continuum of characteristics related to the level of support and challenge that a reader meets in a text (F&P Text Level Gradient™, A–Z). The assessments are available for levels spanning kindergarten through 12th grade and can be used to document reading progress across levels.

Common Assessments at City Year and the Score Types They Offer

	i-Ready	**MAP**	**STAR**	**RI / MI**	**Fountas & Pinnell Benchmark**
Math & ELA	Both	Both	Both	Both	ELA
Test Administration	Adaptive	Adaptive	Adaptive	Adaptive	Fixed Forms
Test Time	30-60 min.	30-60 min.	20 min.	20 min. / 30-40 min.	20-45 min.
Raw Score	No	No	No	No	Yes
Scale Score Range	0-800	100-350	0-1400	0L-2000L / 0Q-1500Q	n/a**
Performance Levels	✓	×	✓	✓	×
Vertical Scale	✓	✓	✓	✓	×
Lexile® / Quantile®	✓	✓	✓	✓	×
Normal Curve Equivalent	×	×	✓	✓	×
Percentile Rank	✓	✓	✓	✓	×
Grade Equivalent	✓	✓	✓	×	×
Student Growth Percentile	×	✓*	✓	×	×

*NWEA MAP calls student growth percentiles "conditional growth percentiles (CGP)."

**The Fountas & Pinnell benchmark system assesses a student's reading level, which ranges from A-Z+.

Interpreting Test Scores

"Although loads of educators refer to "criterion-referenced tests" and "norm-referenced tests," there are, technically, no such creatures. Rather, there are criterion- and norm-referenced interpretations of students' test performances. For example, educators in a school district might have built a test to yield criterion-referenced interpretations, used the test for several years and, in the process, gathered substantial data regarding the performances of district students. As a consequence, the district's educators could build normative tables permitting norm-referenced interpretations of the test which, although born to provide criterion-referenced inferences, can still permit meaningful norm-referenced interpretations." - Popham W. James.

Getting Started

The following scores measure a student's performance on the assessment at a point in time, but don't provide any judgement about what material a student might be expected to know or how the student did as compared to all other students until these score types are combined with performance levels or other scores.

Raw Score – The score that is the sum of points possible for each test question before any statistical processing. Raw scores are not comparable across time or between different administrations of a test. Raw scores cannot be used to measure growth. Also known as the **Observed Score.** We can ignore this score at City Year in most contexts. (Note the DESSA also has a raw score, which City Year does not use.)

Scale Score – A conversion of the raw score onto a scale that is common to all administrations of that test. The scale score takes into account the difficulty level of the specific set of questions, making scores across test administrations comparable. A scale score of 1100 in reading always has the same meaning. Scale scores in different content areas are generally not comparable. Scale scores can only be compared across grades if the assessment provider has developed a single scale across all grades, called a vertical scale.

Lexile® and Quantile® – Lexile® and Quantile® are scale scores with some special features and are often used by City Year. Both were created by MetaMetrics and are licensed for use on multiple assessments including RI/MI, iReady, MAP, and STAR. The brand name refers to both the scale score and the way the scale is developed to be a vertical scale for the level of ability in reading or math.
A Lexile® reader measure is a scale score and represents a person's reading ability on the Lexile® scale. (A Lexile® text measure represents a text's difficulty level on the Lexile® scale and is separate from the test score.) If a reader has a Lexile® measure of 1000L, she is forecasted to comprehend approximately 75 percent of a book with the same Lexile® measure (1000L). Although Lexile® measures should not be linked directly to grade levels, it is possible to describe the Lexile® measures of typical students at various grade levels.
Similarly, there are two Quantile® measures: the Quantile® that represents the math skill level of your student and the Quantile® that describes the difficulty of math skills and concepts they encounter. (There is also a third: lowercase quantile is a statistics term.) The student Quantile® measure attempts to describe skills and concepts your students are ready to learn so you can better personalize math instruction. City Year also uses the Quantile® measure to help place a student to the right range of units within the Do the Math and Do the Math Now! curriculums.

Criterion-referenced Interpretation of Test Results

The Advanced Placement (AP) exams are criterion-referenced exams, as the overall goal is for the grades to reflect an absolute scale of performance which can be compared from year to year.

An interpretation of a test score that is relative to a criterion such as achievement standards or frameworks. In a standards-based achievement framework, criterion-referenced scores provide a measure of a student's knowledge, skills, and abilities as embodied in the achievement standards. Typically, a panel of administrators, academics and teachers decide what content is to be measured. The term "criterion" refers to that content, not a cut-score. In-class assessments written by teachers are often criterion referenced as well (but not standardized). **Achievement Tests** often have criterion referenced interpretations. Here is a portion of performance levels from the Florida Standards Assessments for 2019:

	Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Grade 3	240-284	285-299	300-314	315-329	330-360
Grade 4	251-296	297-310	311-324	325-339	340-372

Performance Levels – Ranges of scale scores for each content area and grade-level, marked by cut-scores and for which specific definitions have been applied. Sometimes performance levels are defined as simply pass-fail, and sometimes they describe multiple levels of performance. In large-scale summative assessment, common performance levels are labelled as below basic, basic, proficient, and advanced. Performance levels may also be called **Performance Standards** or **Benchmarks** or **Placement Tables**.

Cut-scores – A score that marks the beginning of a range of performance. Decisions about the cut score location are often made by content experts using methods that have been established as best practice. Ultimately, the locations of cut scores are judgment calls made by either individuals or groups about the point at which a student has entered the next level of performance. It's theoretically possible, for example, that a given test-development committee, if it had been made up of different individuals with different backgrounds and viewpoints, would have determined different cut scores for a certain test.

Norm-referenced Interpretation of Test Results

The California Achievement Test is a norm-referenced test but was last normed against a national population in 1986, so the norm is not recent.

An interpretation of a test score that identifies whether a student performed better or worse than other students. For example, when a student is in the 75th percentile we mean that, compared to all students in the reference group used to set norms, this student did as well or better than 75 percent. Since the point of reference is how other students performed, the characteristics of the *reference group* used to create the percentile rankings is important for accurate interpretations. If the reference group has different characteristics than the students we are testing, it may not be accurate to compare our students with each other based on their percentiles because we would be concerned that the relationship between student scores would also be different than the one established during norming. It also matters when the norms were established. Overtime there are many factors that can change how students perform relative to each other, so the older the norms, the less accurate they may be.

Grade Equivalent (GE or GLE) – A norm-referenced score typically ranging from 0 to 12 representing how a student's assessment performance compares with that of other students. A fifth-grade student with a GE of 7.6 doesn't necessarily mean that the student is capable of doing seventh-grade math—instead, it indicates that this student's math skills are well above average for the fifth grade.

Normal Curve Equivalent (NCE) – A way of standardizing raw scores received on an assessment onto a 0-100 scale. Normal curve equivalent scores and percentiles are the same at 1, 50, and 99, but not at any other scores. A Student staying at the same NCE from one year can be thought of as achieving statistically normal growth and a gain of 2-3 from one year to the next might be considered exceptional.
(This score is similar to a percentile-rank but preserves the valuable equal-interval properties of a z-score. This is advantageous compared to percentile rank scales, which suffer from the problem that the difference between any two scores is not the same as that between any other two scores.)

Percentile Rank (PR) – The relative standing of a student compared to other students. Useful for evaluating student and group performance on a particular test. For example, an assessment score that is greater than 75% of the scores of people taking the assessment is said to be at the 75th percentile, or 75 is the percentile rank. Percentile ranks cannot be averaged nor used to measure growth. Also often shortened to **Percentile**.

Student Growth Measures

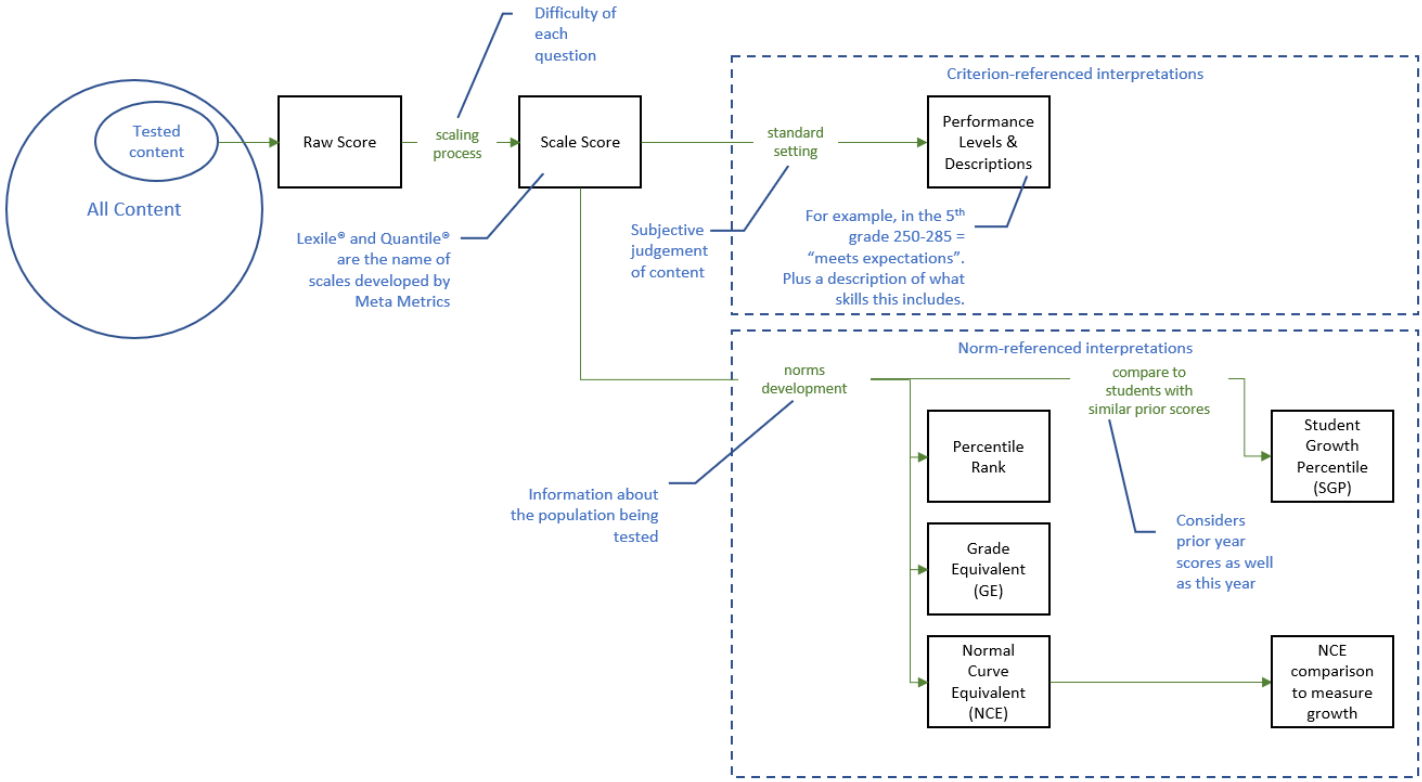
There are three common ways to measure student growth - the choice is dependent on purpose and test design.

Student Growth Percentile (SGP) – A measure of student academic growth, as compared to similar students. Typically, "similar students" are defined as students in the same grade, with similar scores from last year and similar starting scores. A student with an SGP of 30 scored higher than 30 percent of students with similar score histories.

Student Growth Along a Continuous Scale (Vertical Scale) – A way to measure student growth in a content area, *within and across grades*, using scale scores achieved over continuous administration of progressively more difficult tests. This kind of growth measure is made possible through the development of a single test scale across all grades. The difficulty of these tests generally increases as students progress through grade level material, allowing us to determine growth simply by comparing students' scale scores as they progress through the tests.

Student Growth Using NCEs – NCEs provide a way to measure student growth across grades when the tests are not on a vertical scale. In this case, students' scale scores are converted to their normal curve equivalents and growth is determined by evaluating changes relative to what is consider *normal* growth. Normal growth is typically thought of as achieving the same NCE in successive grades, because the examinee falls in the same position, relative to other students, over progressively more difficult content.

Diagram of Types of Test Results



A Note on Standards - The word *standards* can be confusing. It helps to be aware that this term is commonly used in three different testing contexts. First, *standards* refer to "Achievement Standards" which represent the content domain and learning expectations. Achievement tests are most often built in alignment with these standards. Second, the term *Standards* is often used interchangeably with the terms **cut-score** or **benchmark**, which as described above, refer to the points on a test scale that define different performance levels. This use of the term comes from the name of the process used to determined cut scores or benchmarks, which is called "Standard Setting." Finally, a *standard* in testing can refer to any component of the professional standards (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education's, 2014) that are used by test developers to build tests that produce valid and reliable scores.

Test Administration

Tests can be either paper-based (PBT) or computer-based (CBT). (When both types of administration occur for one set of items, we are concerned about whether there is evidence that scores are comparable across the two modes of test administration.)

Fixed-form – Traditional tests. A test where every student answers every question in that test administration. Can be paper-based or computer-based.

Adaptive – A computerized assessment that adjusts what questions are asked based on the responses to prior questions. Usually provides instantaneous results. In an item level adaptive test, each new item is assigned one at a time based on the students performance so far. In a multistage adaptive test, students see items in small sets. One difference

is that, in item level adaptive tests, students usually cannot go back to any questions once they have been answered. In a multistage adaptive test, students can return to items within each set.

Testing Conditions – The physical and mental conditions under which a student takes a test are important for the results to be valid. Generally, the most important factors are low stress, adequate space, good classroom management to prevent disruptions and sufficient time. For example, if a student does not see the value in an assessment, they may "click-through" an adaptive computer assessment, choosing the first option for each question. Note that students with an IEP or 504 plan are often provided with modified conditions such as extra time or a private space. Test providers should provide a test guide that lists conditions, and there is usually a test coordinator in each school.

More nerdy stuff

More information on growth models

https://scholar.harvard.edu/files/andrewho/files/a_pracitioners_guide_to_growth_models.pdf
https://www.nciea.org/sites/default/files/inline-files/Understanding-selecting-implementing-growth-measures_5-27-16_0.pdf

Other less common scores

Standard Score – A conversion of the raw score onto a scale that fits a normal curve. Like a scale score, a standard score is derived from raw scores and transformed. Standard scores indicate how far above or below the average (the "mean") an individual score falls using a common scale, such as one with an "average" of 100. Standard scores also take "variance" into account, or the degree to which scores typically will deviate from the average score. Standard scores can be used to compare individuals from different grades or age groups. Also called **Z-Score** when the mean is 0 and standard deviation of 1. **T-Score** is a standard score Z shifted and scaled to have a mean of 50 and a standard deviation of 10. (In addition to raw score and percentile, the DESSA reports a T-Score.)

Stanine – A norm-referenced score with a scale between 1 and 9. Has a mean of 5 and a standard deviation of 2. Represents equal units of achievement and can be averaged.

References

Developed in collaboration with the Center for Assessment and City Year

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Marion, S., D'Brot, J. & Martineau, J. (2019). *An introduction to assessment types and uses*. Center for Assessment Policy Brief. Dover: Center for Assessment.

https://www.edglossary.org/

https://en.wikipedia.org/wiki/Educational_assessment

https://www.renaissance.com/2018/07/11/blog-criterion-referenced-tests-norm-referenced-tests/

Popham, W. James. 2011. Classroom Assessment: What Teachers Need to Know (Sixth Edition). Boston: Pearson Education, Inc. pp. 46-8.

Laming, D. R. J. (2003). Human judgment: The eye of the beholder. Australia: Thomson Learning.

Also various assessment provider materials