So What Does it All Mean? A Primer on Testing Terminology

The education world is increasingly focused on the use of data in general, and testing data in particular, to drive the decision-making process about individual students, curriculum, programs, and, most recently, teacher and principal effectiveness. What we have paid less attention to, unfortunately, is educating our community—students, teachers, parents, and other stakeholders—as to what different test scores mean. This primer is designed to help explain, demystify, and uncover some of the more common terms and practices surrounding testing and test scores.

A. Testing Terminology

Criterion- vs. Norm-Referenced

Tests are either **Criterion-Referenced** or **Norm-Referenced.** Criterion-referenced tests measure the degree to which an individual has mastered the expected content, often including all the expected content at a single level of learning. These tests are only capable of measuring how well a child has done on the level it is written to measure (most often a single grade level, for part or all of a year). The New York State Testing Program, which includes the grades 3-8 Math and ELA tests, and Regents exams, is exclusively made up of criterion-referenced tests.

Alternatively, norm-referenced tests compare one individual to others who took the same test. Content on norm-referenced tests typically includes only questions that are good at differentiating between various levels of student knowledge. Individual achievement tests (OLSAT, AIMSWeb, SAT, ACT) are always norm-referenced.

National vs. Local Norms

Some norm-referenced tests return two sets of results: scores based on **national norms**, and scores based on **local norms**. National norms are based on the group of students of the same grade who were tested to establish the test's results, during test development. If the test is well-designed, the children in this normalization population should include a cross-section of gender, race, income, urban-suburban-rural schools, etc. Local norms are scores generated based on the specific students in this school or district, in this grade, taking this test on this test date. With this kind of scoring, you see not only how your child compares to kids across the nation, but also to kids in your local district and classrooms.

B. Statistics "101"

Mean is a measure of the average score achieved by the population taking the test.

Percent correct is generally used in the classroom, not on achievement tests. Percent correct is easy to understand. 90% correct means that the child got 9 of 10, or 90 of 100, or a similar number of questions correct. It is important to distinguish *percent* from *percentile*.

Scale Score is a method of converting a student's raw score on a test to a common scale that allows for numerical comparison between students. Scale scores are primarily used to compare test scores over time, such as measuring year-to-year growth of individual students or groups of students in a content area. The NYS Testing Program reports student results as scale scores, and performance levels.

Performance Levels are used to broadly describe a range of performance. New York State uses four performance levels, as described below. Each has a corresponding range of scale scores. In

essence, performance level describes the qualitative achievement of the student relative to the standards, while scale score describes achievement more specifically.

- **Level 1- Below Standard**. Student performance does not demonstrate an understanding of the English language arts knowledge and skills expected at this grade level.
- **Level 2- Meets Basic Standard**. Student performance demonstrates a partial understanding of the knowledge and skills expected at this grade level.
- **Level 3- Meets Proficiency Standard**. Student performance demonstrates an understanding of the knowledge and skills expected at this grade level.
- **Level 4- Exceeds Proficiency Standard**. Student performance demonstrates a thorough understanding of the knowledge and skills expected at this grade level.

Standard Performance Index is an estimate of the number of questions a child would have answered correctly if there were 100 items in that cluster, or standard, of content. The child's performance on each standard is compared with the SPI target range. A student scoring within the target range demonstrates the expected understanding of that content standard. The target ranges vary across standards because some standards may contain more difficult items than others.

Percentile describes the percentage of subjects taking the test that scored above or below the individual child. For example, 50th percentile mean half of the students scored below, and half scored above. It does NOT mean the student answered half of the questions wrong, and half right. This way of reporting results is not always helpful. For example, imagine a test in which the teacher *expects* many or most of the students to know most or all of the material. In such a case, the percentile scores for top-scoring would seem skewed. So, if 10 percent of the tested students answered all the questions correct, they would each be reported as being in the 90th percentile, despite the fact that there was no way to achieve a higher percent correct (100% is the top!). Conversely, a student answering only one question incorrectly would put her *below* the 90th percentile. That is where the raw score can help make some sense. However, if properly and appropriately applied, the difference in achievement between a child in the 50th and 59th percentile is fairly small, but the difference in level between a child in the 90th and 99th percentile is very large.

Stanine is another way of describing percentile. Stanine divides the percentiles into 9 divisions. The 4, 5 and 6th stanine are considered average, while the 7th and 8th stanine are considered above average, and the 9th stanine is considered well above average. The percentage of test scores in each stanine is as follows:

Stanine	Percent of Scores	Percentiles
1	4	0-3
2	7	4-10
3	12	11-22
4	17	23-39
5	20	40-59
6	17	60-76
7	12	77-88
8	7	89-95
9	4	96+

Raw score is simply the number of questions answered correctly out of the number of questions available. Comparing a raw score to a percentile can be helpful in understanding if percentile was appropriately applied (see above).

Age- and grade-level equivalents describe the age or grade of the average child receiving the same score as this child.

Standard deviation is a statistical measure of spread. One standard deviation is the range which includes 65% of all scores, two standard deviations includes 95% of all scores.

C. Intelligence and Achievement Tests

While there are probably many, many different categories of test, perhaps the two most common in formal education are the intelligence test, and the achievement test. **Intelligence tests** tell us how capable a person is of learning. These tests do often use acquired knowledge as one indicator of intelligence among many. **Achievement tests** tell us how much of something taught has been learned. Both intelligence and achievement tests can be standardized, meaning they are scored in a consistent, predetermined, standard, way.

Within these two major types of tests, there are two major subsets of tests: group tests, and individual tests. Group tests are usually written and are given in silence to a large group of children. Individual tests are conducted mostly verbally in a one on one situation with tester and the subject.

Group intelligence tests are commonly used as screening measures, to see if the child should move to a more comprehensive assessment. Group tests are generally normed on populations of all children. BCSD administers the OLSAT, a group intelligence, or school ability, test. It tests students likely ability to cope with typical school learning tasks.

Individual intelligence tests are administered one on one, and include tests like the Wechsler tests, Wechsler Intelligence Scale for Children (WISC-IV) and Wechsler Preschool and Primary Scale of Intelligence (WPPSI-III).

Grade-level or group achievement tests are criterion-referenced, so they contain questions covering just about every aspect of the curriculum at that grade level. Grade-level achievement tests are normed for no more than a single grade level, and can only determine if the child is at, below, or above grade level. All state-mandated grade level achievement tests are group achievement tests.

Curriculum Based Assessments (CBAs) are specific to the curriculum being taught in a school. Examples include quizzes, unit tests, interim benchmark assessments, and mid-term or final exams for a course of study. CBA assessment is directly related to the local curriculum, and to available intervention and instructional planning. CBAs are considered reliable and valid.

References

http://www.p12.nysed.gov/irs/ela-math/, accessed September 2, 2011

http://www.hoagiesgifted.org/tests_tell_us.htm, accessed September 2, 2011.

http://jcacardinals.org/Test%20score%20Explanation.pdf accessed September 2, 2011