Section 5
WHEN to Use Which Types of Test Scores
Test Scores and What They Mean

- Test Scores
- Norm-referenced vs. Criterion-referenced
- Helpful Tips in Interpreting Your Scores to Others
- Professional Development

Assessment and evaluation generate lots of questions from parents, board members, administrators, teachers, the community, and even students. The major education associations believe that teachers should possess strong assessment competencies: "Teachers should be skilled in using assessment results when making decisions about individual students, planning, teaching, developing curriculum, and improving schools." (AFT, NCME, & NEA, 1990). With high stakes testing, this requirement is even greater today than previously. Yet, many teachers and administrators receive little training. Thus, we hope that this section of the Toolkit will help teachers and administrators understand some important aspects of test scores and how to interpret them.

Test Scores

Test scores form the heart and soul of most program evaluations. Everyone has taken tests, many have given tests, and we generally have a sense of whether the scores look good or bad. However, it is important to understand the different kinds of scores that might go into the database for subsequent analysis so we can make sure to have the right kinds of scores.

Let's be clear from the beginning that "scores" don't necessarily have to come from tests. Scores can come from questionnaires and surveys or just counting people. Anything that attaches a number to what you're doing in a program—the numbers of kids served, the numbers per language groups, parents' average years of education, responses on a scale of 1 to 5 on an observation form or survey—those are all scores and can be analyzed.

Raw Scores

Let's dispense with raw scores right away. A raw score is simply the number of correct responses on a test. Unless you know what the test is measuring and how many possible points it had, a raw score doesn't mean much. Raw scores may be useful at the local level if everyone knows the test and what it covers, how many possible right answers there are, and how students typically do. However, most test scores that are important to program evaluation are transformations of raw scores and take the form of norm-referenced or criterion-referenced scores.

Norm-Referenced vs. Criterion-Referenced

You are already familiar with both kinds of scores, even if you haven't thought about them in those terms. In short, a test score is **norm-referenced** if it gives you a number that tells whether a student is roughly average in relation to most similar students of his or her age or grade, the student is relatively above average, or the student is relatively below average. Norm-referenced scores compare people with each other. A test score is **criterion-referenced** if it is compared to a preset standard or level of achievement.

> **Norm-referenced**: measures broad skill areas, then ranks students with respect to how others (norm group) performed on the same test. Students' scores are reported in percentiles, stanines, or normal curve equivalents. It is impossible for ALL students to score above 50th percentile. Norms are established so that 25% will score in bottom 25th percentile, 50% below 50th percentile, 75% below 75th percentile, and so on.
>
> **Criterion-referenced**: determine whether students have achieved certain defined skills. An individual is compared with a preset standard for expected achievement. The performance of other students is not important to the interpretation of a particular student's score. On criterion-referenced tests, it is possible for ALL students to achieve the minimum achievement expectation. Criterion-referenced scores provide a number that tells more or less how well someone performs on a task, regardless of how anyone else does: the student is very good at it, the student has some command of it, or the student has a long way to go before being good at it. Most of the state tests are criterion-referenced; that is, there are state standards that define what knowledge a student should have at each grade level. These tests are designed to measure whether each student has learned the expected knowledge.

Examples of Norm-Referenced Scores and How to Interpret Them

The following kinds of scores are the most common norm-referenced scores used by educators. As stated in the beginning of this section, they don't tell how well a student does something, just whether the student is above or below average.

- **Percentiles** tell what percentage of students at the same grade level got lower scores. For example, if a student has a percentile score of 45, he or she did better than 45% of students at his or her grade level. The 50th percentile is the statistical average, so a student with a score at the 45th percentile is a little below the statistical average but not very far below.

- **Normal Curve Equivalents (NCEs)** are a conversion of percentiles. Percentiles can not have statistical analyses such as averaging done with them, but NCEs can. (See the Appendix for a conversion table between Percentiles and NCEs). The concept for NCEs is the same as for percentiles; that is, a score of 50 NCEs is statistically average. An NCE of 45 is somewhat lower than that, but certainly higher than 20 or 25. Similarly, an NCE of 75 or 80 is well above the statistical average. An NCE score indicates where a student stands in relation to grade-mates, but little else. Student score reports usually come back in percentiles. However, if you have the choice, enter the NCE scores rather than the percentile because you can do more with them.

Examples of Criterion-Referenced Scores and How to Interpret Them

Criterion-referenced scores make a statement about how well a student performs, regardless of how other students perform. Common examples include:

- **Grade Equivalents** represent the extent to which a student can read material typical of a student at a certain grade. For example, a student with a reading score of 5.2 can read the material typical of the second month of fifth grade. However, those scores become very unreliable as they get further from the student's actual grade level. For example, if a fifth-grade student gets a reading score of 9.3, that student reads very well, but he or she cannot necessarily effectively read the same material that a ninth-grade student can. For one thing, the fifth-grader probably does not have the experiences associated with ninth-grade material.
- **Scale Scores** are usually three-digit scores that have been converted from raw scores, where high scores indicate the ability to do difficult work as measured by the test, and low scores indicate the ability to do only easier work. Scale scores have become more common as states have developed tests to examine whether students are reaching grade-level expectations defined by state content standards. Scale scores mean different things for different tests, so it's important to consult the technical manuals to understand them. (Different tests use different scales. One might report scores between 180 and 250, for example, another in the 200's, and another in the 300's. It doesn't matter. A score of 200 on one test might mean something totally different from 200 on another test. You just need to familiarize yourself with the meanings of the scores of the test you are using.) Also, you need to be clear on what type of scale score your state test uses:
  - **Scale Scores —** Some states (e.g., California) use scale scores in which the score is dependent on the grade level, and scores cannot be used to examine growth over time. For example, in California, the scale scores for each grade and subject area range between 150 (low) to 600 (high). In California, then, for all CSTs, the minimum scale score required to achieve the proficient level is 350, which is the goal for all students. A student who scored 350 in grade 3 and again in grade 4 would remain at the proficient level.
  - **Vertical Scale Scores —** Other states (e.g., Oregon, Washington) use vertical scale scores; that is, the scale scores are independent of grade level, and can be compared for growth over time. Thus, a score of 350 represents the same difficulty, whether the student is a second-grader or an eighth-grader. However, the standards may be different for each grade level. Perhaps the typical second-grader is only expected to do the kind of work represented by a score of 350. The eighth-grader may be expected to do the kind of work represented by a score of 480. Thus, 350 may be a satisfactory score for a younger student but not a satisfactory score for an older student.
- **Performance Categories** are commonly used in education. They are also called "levels." Language proficiency tests often yield five levels, from beginner (level 1) to

proficient (level 5). In the No Child Left Behind (NCLB) context, academic achievement tests also yield performance categories or levels: advanced, proficient, basic, below basic, etc. Performance categories, or levels, are actually specified ranges of another kind of score, such as percentiles or scale scores. They group the more fine-tuned information of those kinds of scores. The advantage is that they communicate something about what a student can do. The disadvantage is that they lose important information. For example, let's say level 1 consists of scores 1-20, level 2 consists of scores 21-40, etc. Two students who have scores of 22 and 38 are considered in the same level, while students who are very close, say 19 and 21, are considered to be at different levels.

HELPFUL TIPS IN INTERPRETING YOUR SCORES TO OTHERS

Here is another important difference between norm-referenced scores and criterion-referenced scores.

If a student’s **norm-referenced score** (NCE, percentile, stanine) does not go up over time, that does not mean the student has not learned more. It merely means the student has maintained his or her position relative to other students. Think of a footrace. The runners may finish the race in the same order that they were halfway through the race. They all made progress. It’s just that the runner in the middle didn’t catch up to the runners in front.

In contrast, with **criterion-referenced scores**, the expectation is that students with lower scores will go up as they learn more. We want students who scored “below basic” to move up to “basic,” then “proficient,” then maybe even “advanced.” We want students who started at Level 1 in English to move up to Level 2, then Level 3, etc.

It is important to understand the kinds of scores that are being used in the evaluation in order to interpret and explain them.

Section 6 of this Toolkit provides detailed information on data analysis. Different analyses are appropriate for different kinds of scores, as that section explains. It may be useful to study that section and then return to this discussion afterwards.

Note. As you go through the Toolkit, if you forget what some of these terms mean, you can look them up in the Glossary.

Professional Development

"Assessment competencies are an essential part of teaching and good teaching cannot exist without good student assessment" (AFT, NCME, & NEA, 1990)

The major education associations strongly believe that teachers should possess strong assessment competencies, including the following:

- Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and improving schools
- Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators (AFT, NCME, & NEA, 1990)

Developing these competencies may require training at the school site. Training may be required in understanding achievement test scores—how to interpret them and how to use them as one indicator to diagnose students' strengths and weaknesses. Once teachers have this knowledge, they have no difficulty in communicating with parents and others about student performance. Other training should include assessment instruments that are developed at the district or school site level or by another educator to assess particular student or classroom characteristics.

For example, the authors of this Toolkit have found that when teachers were not well trained in the assessment rubrics, there were reliability problems:

1. Some teachers were much too liberal and others too conservative in their ratings.
2. Expectations for language minority and language majority students affected teacher ratings in ways that favored language majority students. Because individuals in the US typically do not expect a native English speaker to speak in Spanish, the scores of these students tended to be higher in Spanish than the scores in English of the language minority students, who were expected to learn to speak English.

Finally, we found that some teachers collected the portfolio data, but never used it. While it may seem obvious to some teachers how to use portfolio data, training may be needed to assist teachers in using this rich source of information for making instructional decisions about students.