

Understanding Why STAR Test Scores Fluctuate

Scores on any test may decline from one administration to the next for a variety of reasons, some of which are easy to understand, while others are difficult. Three of the reasons scores decline are the following:

1. Measurement error, which is present in any test
2. Fluctuations in student performance
3. A statistical phenomenon called "regression to the mean"

Measurement Error

Educational tests are all imperfect; as a consequence, every educational test score contains some degree of error. As a result, if a group of students were to take a test twice, some of their scores would be expected to decline, even if their skills have improved during the interval between tests. To determine whether a student's score has declined more than would be expected, one must compare the amount of the score decline to a statistic called the "standard error of measurement" or "SEM." The SEMs for STAR tests vary somewhat by grade and by score level; their typical magnitudes are tabulated in the technical manuals for each of the STAR tests. As a rule of thumb, a score decline should not be considered meaningful unless it is more than about 1.5 times the SEM. For example, imagine that a student obtained a scale score of 600 on STAR Math™ in September, and 550 when she took it again in October; if the SEM for STAR Math is 40 points, the 50-point decline should not be considered significant since it is less than 1.5 times 40 (the SEM), or 60 points.

If there has been no learning at all between the first and second test administration, about half of all scores will increase, and the other half will decrease, solely because of measurement error. Although individual students' scores may fluctuate substantially, the average score of all students will be nearly identical on both the first and second tests.

If there HAS been learning growth between test administrations, some scores may still decline because of measurement error, but the percentage of students with declining scores should be less than half, and more than half of all scores will normally increase. The longer the interval of time between successive test administrations, the smaller the proportion of score declines will be.

When some scores decline, there is a natural tendency to distrust the validity of the test. This is inappropriate, however, if measurement error alone would explain the decline -- as is often the case in educational testing. As a check on the validity of a test where there are some score declines, a teacher should compare the average scores and the spread of scores on the first and second test administrations. Has the group's average and spread of scores increased? Is the increase close to what is expected, given the time interval between tests and the typical amount of growth expected over that interval? If the answer to each question is "yes", there is no reason to be concerned about the validity of the test scores.

Fluctuations in Student Performance

A test score is a measure of a student's performance on a given day. Human performance is not always consistent from one occasion to another. A student may perform at her best during one administration of a test, and somewhat less than her best on another occasion. These fluctuations may be related to such things as illness, distractions, anxiety, attention or motivation. We are not surprised by such fluctuations when they affect physical performance -- such as running speed or jumping distance in athletic contests. Similarly, we should not be surprised that inconsistent test performance may result in declining scores on occasion.

Regression to the Mean

This is an inevitable phenomenon that occurs whenever there are multiple observations of an attribute that is not perfectly consistent from one occasion to another. Educational test scores are a classic instance of that. When students are measured twice, there is a tendency for those with the highest scores on the first test to score somewhat closer to average (and therefore lower) on the second test. The reverse happens at the low end of the scale: students who scored lowest on the first test tend to score somewhat higher (and thus closer to average) on the second test. There is a tendency for all scores to regress to the mean but this effect is more highly pronounced for the highest and the lowest scores. If the time interval between the tests is fairly short -- a few days or even two or three months -- there may be a substantial number of students whose scores decline even though their skills have improved. Regression of scores to the average occurs since the degree of correlation or score order between the two sets of scores is less than perfect-which is typically the case.

What to Do When Scores Decline

The first thing to do is to compare the average scores on different occasions, for the class or a larger group. If the average scores are increasing over time, then one or more of the three causes above may fully explain those score declines that have occurred. To determine that, compare each individual score decline to the SEM for the test; if an individual student's score decline is less than 1.5 times the SEM, there is little reason for concern. Chances are that students with declining scores will show score increases the next time the test is given.

In cases where a student's score has declined by some larger multiple of the SEM, the teacher is wise to investigate more deeply. Was there something in the student's personal life that may have depressed their test performance on that day? Does the student seem to be responding well to instruction? If the teacher is concerned about large individual score declines, one possibility is to administer the test again, under closely monitored conditions. Another is to seek advice, perhaps from a staff member who is well versed in the technical aspects of assessment, or even from the test publisher's technical support department.

In conclusion, we should expect some scores to decline when a test is administered more than once over a short interval. The size and frequency of the score declines should be compared against both expected score growth and the published magnitude of the SEM before concluding that a test score is not valid. Having done that, any score declines that are too large or too many to be explained by those factors can be investigated further, with the help of resource persons available locally, or by calling on the test publisher's technical support staff.