Standard Error of Measurement: A Concept That Every Gifted Education Specialist Must Understand

By Karen L. Westberg
University of St. Thomas

Imagine that a James, a Grade 2 child, took a group-administered aptitude test in school, the *Cognitive Ability Test (CogAt),* as part of the screening process for receiving gifted education services. And, let's say James received a composite score of 122 on this test. Does this score represent his "true score", a score he would obtain again if repeatedly tested on the same test? Actually, the likelihood of this happening (obtaining the same exact score) is quite rare.

Using Standard Error of Measurement to Interpret A Test Result

To illustrate why obtaining the same score when hypothetically being tested again is rare, let me explain what we can say about James' obtained score of 122. We can use the reported "standard error of measurement" score to estimate a range of scores in which James' "true score" lies. In the administration and technical manual accompanying the CogAt test, the publisher (Riverside Publishing) reports that the standard error of measurement on the CogAt (for the verbal, quantitative, and nonverbal subtests) is 3 points (Note, the reported standard error of measurement varies on tests…another aptitude test would likely report a different standard error of measurement.) With this piece of information (3 points), we can now estimate the range of scores in which James' "true score" lies, and share the probability for obtaining a score within this range.

To arrive at this estimation, we use the probabilities on a normal curve figure (also known as the bell-shaped curve or normal distribution curve) to estimate the range in which James' true score lies. If we plot a score of 122 in the center of a normal curve figure, we would add three points (the reported standard error of measurement on the CogAt) to the first standard deviation mark above the mean, labeling it 125 points, and we subtract three points from 122, making it 119 points, and place a 119 on the first standard deviation mark below the mean. The area of the normal curve between one standard deviation marker below the mean (in this example, 119) to one standard deviation marker above the mean (in this example, 125) is 68%. Using the probability associated with those two places on the normal curve figure, which is also known as a normal probability curve, we can now say that there is a 68% probability that James' true score lies between 119 and 125 points. In other words, when tested again, there is a 68% likelihood that his score would be between 119 and 125. Stated another way, we could also say with a 68% probability that this is the range (119 to 125) in which his "true score" lies.

Maybe you don't like those odds, a 68% likelihood, and you would like to be more confident about the range for his "true score", or score when hypothetically tested again? If so, we would now add three points above 125 points to the second standard deviation marker above the mean, placing a score of 128 as this marker, and we would subtract 3 points from 119, placing a 116 score at the second standard deviation marker below the mean. When doing this, we now estimate that there is a 95% probability that James' true score lies between 116 and 128. Obviously, as the probability increases from 68% to 95%, the range or band of potential scores increases. In other words, if we want to be more confident about the range in which a score would occur when someone is hypothetically tested again on the same test, the range of scores would be larger. Where does the 68% value or the 95% value come from? The normal curve figure below illustrates the portions or the curve (and probabilities) associated with standard deviation markers or signposts above and below the mean. These values are the same or standard for all normal curves.

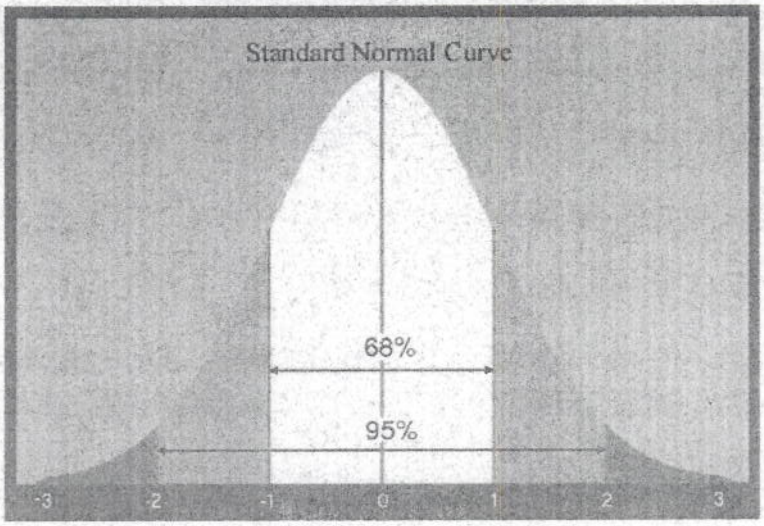

Figure 1. Standard Normal Curve

You might be asking yourself now why someone would not get the exact or nearly exact same score when tested again on a particular test. It occurs because no tests are perfect. There is imprecision in all psychological or educational measurements. "Standard error of measurement" is an estimate of the error to use when interpreting an individual's test score. The calculation of this value is beyond our discussion here, but realize it involves the use of the test's reliability coefficient that was obtained by the test developers when they conducted research on the instrument.

Implications of Standard Error of Measurement for G/T Identification

What are the implications of this issue if we use test scores when identifying students for gifted education services? If you use cut-off test scores as your only tool for determining placement decisions, you might be eliminating a child from placement because he or she might obtain a different test score on another day, and it has nothing to do with what a child knows or doesn't know—it has to do with the imprecision of tests. This is why we increasingly are seeing reports of test results that provide a band of scores, with instructions about how a child's score was found within that rangeor band. The greatest implication of this is that a single test score should not be used as a sole criterion to determine whether or not a child will receive services, and this is why we recommend using multiple measures for identification, such using teacher judgment measures along with test information. Using a single test score is simply an indefensible practice.

Where can you find information about the "standard error of measurement" on tests? The publishers of commercial tests should report this information in their technical manuals. If you don't find it reported there, you can usually find it on publishers' websites. One of the things you will notice when examining reported "standard errors of measurement" is that they vary according to the test. The reported standard error of measurement scores are not the same for all published tests. For example, the standard error of measurement on the Non-verbal subtest of the Stanford-Binet, 5th ed., is 3.9 points; and the average standard error of measurement on the Naglieri Nonverbal Ability Test (NNAT) is 6.1 points. Of course, several factors (cost, administration time, validity issues) are considered when selecting which ability or achievement test to use, but "all things being equal", you would want to use a test with a small standard error of measurement.

When you find the reported "standard errors of measurement" in a technical manual, you will learn that a test's standard error of measurement is not the same across a continuum of scores; it often varies slightly depending on the age or grade level for the test (a test's standard error of measurement might be 3.2 for Grade 2 and 3.0 for Grade 3) and varies for different obtained scores; namely, the standard error of measurement is sometimes larger for extreme scores, low or high. For example, the standard error of measurement on the NNAT is not 6 points for all obtained scores. For students obtaining very high scores, the standard error of measurement on the NNAT is larger.

Standard error of measurement is something that all gifted education specialists (and, I would argue, "all educators") should understand. Armed with this knowledge, our decision making will be more defensible, and students will be better served.