

## FITNESSGRAM Training

### Validity and Reliability

**Validity** is an appraisal of whether a test measures what it's supposed to measure. Usually it's expressed as a statistic called a *correlation coefficient*, which quantifies how strong the relationship is between test results and the most direct measurement of the fitness component that's being tested. A coefficient runs anywhere between -1.00 and +1.00, with an acceptable level being .80 or higher. (The larger the coefficient, the stronger the connection, whether positive or negative.)

**Reliability** is a measurement of whether the same student gets about the same score when performing the same test repeatedly. Physical fitness tests usually show high reliability, especially when students put out a maximal effort. Reliability may drop if students don't give their best effort, are not sure how to perform the test task, or are feeling unhappy, tired, or stressed.

When a criterion-referenced test, like FITNESSGRAM, is developed, both the validity and the reliability of the standard and the validity and the reliability of the criterion must be determined. For instance, the PACER test has to be shown to validly and reliably measure aerobic capacity; the standard set for that test must also be shown to validly and reliably relate that amount of aerobic capacity to the health outcome desired (such as reduction of heart disease risk).

To set a valid criterion-referenced standard, test developers must have both scientific knowledge and measurement expertise. They set standards based on expert judgment, knowledge of the score distribution of the field test and the criterion test within the population, and the relationship between the field test and the criterion measure. The standard represents the level of risk for the health outcome related to the fitness component being tested. It may be a single score or a range of scores. FITNESSGRAM uses a range of scores called the Healthy Fitness Zone (or HFZ) for each test item.

For a criterion-referenced test to be *valid*, all those who pass the criterion measure should also pass the criterion cut-off score on the field test. Conversely, all those who fail the criterion measure should also fail the criterion cut-off score. The test should consistently classify people into one classification or another. However, sometimes it will not, since no field test is perfectly valid and measurement errors can occur.

For a criterion-referenced test to be *reliable*, it also should classify people as meeting or not meeting the standard consistently when the test is repeated. If the test is given one day and then again the next, the student should be classified as either meeting or failing to meet the standard both days if the test is reliable. Consistent results should also occur if two different people administer the test (called *inter-rater reliability*) or if the same person administers the test both times (called *intra-rater reliability*).

To determine the validity and reliability of a criterion-referenced test, test developers use four types of statistical analyses to check on how reliable the results are and how closely the test results relate to the underlying criterion, making sure the relationship isn't just based on chance. They adjust the cut-off score or score range until it best classifies students. Then they compare the cut-off score or range of scores across analyses. If the analyses agree on the cut-off score or range of scores, it's likely that it is valid and

reliable and should be used as a standard. If the analyses don't agree, a problem may exist with either the test or the criterion, or perhaps the quality of the data used for analysis.

### **Examples of FITNESSGRAM Criterion-Referenced Standards**

The criterion for a criterion-referenced fitness test standard should be based on scientific evidence. Let's look at how this was done for the body composition test in FITNESSGRAM.

Obesity has been shown to be a risk factor for cardiovascular and other chronic diseases, and extremely low body fat can be related to eating disorders, so there are health reasons for tracking body composition. Two field tests commonly used to measure levels of body fat are skinfold thicknesses and Body Mass Index, and these are included in FITNESSGRAM. Once we have the measurements obtained using these tests, we need to have something with which to compare them. We need to know at what percentages of body fat a person's health starts being adversely affected, whether the percentage is low or high.

How do we find out? Comparison with national norms won't help us decide, as it's generally acknowledged that the U.S. population has been becoming more obese over time. So we need to look at epidemiological studies. By studying large populations and measuring their level of body fat using a criterion measure of underwater weighing (the most reliable measurement method), researchers can determine how many people at what levels of fatness have health problems or risks. For instance, they may find that a percentage of body fat between 16.4% and 25% is desirable for 16-year-old girls because studies have shown that girls of this age whose body fat falls within this range have decreased health risks and/or better health.

Once we know what ranges of percentages of body fat are desirable for boys and girls of various ages, we can then convert the test scores from measuring skinfolds or the Body Mass Index to body fat percentages and set test standards.

A slightly different procedure was used to set standards for the 1-Mile Run Test. Here three criteria were used: VO<sub>2</sub> max, running efficiency, and speed of running. All three values were used to calculate the standards for this test. (See chapter 5 in your manual for more details on this, or go to [www.FITNESSGRAM.net](http://www.FITNESSGRAM.net) and click on "Reference Guide.")