



LEARNING POINT

Criterion- and norm-referenced score reporting: What is the difference?

Scores on educational tests can be reported in two ways: criterion-referenced and norm-referenced. These two notions describe the context in which a student's score on a test can be interpreted. Understanding the difference between these two frames of reference is important, not only for the interpretation of test scores, but also for creation or selection of tests for specific purposes. Given a particular desired use of an assessment, one frame of reference might be more appropriate than the other.

This document will cover some aspects of norm- and criterion-referenced

be as simple as adding up the number of items answered correctly or the level attained on a rubric-scored item. More complicated scoring rules include differential weighting of items or test sections. A raw score is the most basic form of score on an educational test.

Interpreting a raw score on its own, however, is difficult to do. Suppose for example, a parent comes to you and states that their child just took a test and "got a 35 on it." They ask you if that was a good score. How can you answer this question? Clearly more information is needed. The first question you would probably ask is, "What test was this on?" This knowledge would provide information about the scale that is being reported. We would interpret the 35 very differently if it were a score on a final exam based on 100 points (a percentage), on the ACT (out of 36), or on the SAT (off the scale). Understanding the scale on which a score is presented is the first step to interpreting the score.

Another basic type of raw score reporting is the percentage of test items that the student answered correctly. Thus, if a student correctly answered 38 out of 50 questions, we can say that the student answered 76% of the items on the test correctly. But, like raw scores, is 76% a good score or not? Well, if this was a very easy test (that is, most students answered a

higher percentage of the items correctly), then a score of 76% is not very good. Conversely, if most students answered a smaller percentage of items correctly, then the score of 76% is quite good. As with raw scores, more information is needed in order to interpret percentage scores.

A caution is in order when thinking about percentage scores. As illustrated in the previous paragraph, using the same passing level for all tests, such as "students need to score above 70% in order to pass the test," does not, necessarily, make the passing results of different tests comparable. The passing score needs to take into account the difficulty of the overall assessment. Answering 70% of the items correctly on a very easy test doesn't demonstrate the same level of knowledge or mastery as does answering 70% of the items correctly on a very difficult test.

Criterion-referenced scores

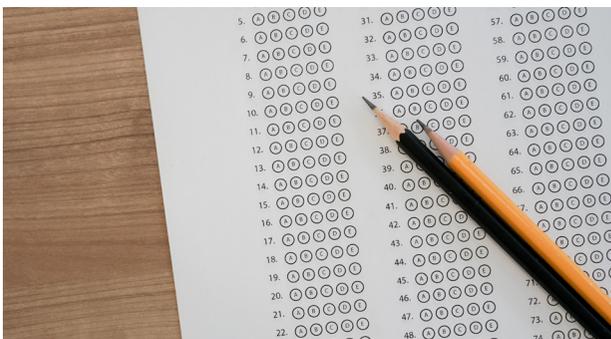
A criterion-referenced test score compares a student's raw score to a predetermined standard based on the content of the assessment. A conversion of a raw score to a percentage based on the total points possible is a familiar criterion frame of reference to those who have experienced educational testing.

Another way to provide a criterion-referenced score is in terms of mastery. Many educational tests now report student scores in terms of levels

scores. We will start by discussing raw scores as those are the scores that norm- and criterion-referenced scores are derived from.

Raw scores

The most basic type of score on a test is the raw score. A raw score is assigned to the results of a test based on scoring rules. The scoring rules could



TO LEARN MORE

Psychometric theory, 3rd edition. Jum C. Nunnally, J. & Ira Bernstein. (MacGraw-Hill, 1994)

Glossary of Education Reform Great Schools Partnership

Norm-referenced test – www.edglossary.org/norm-referenced-test

Criterion-referenced test – www.edglossary.org/criterion-referenced-test

Measurement and evaluation: criterion- versus norm-referenced testing (Table)
www.edpsycinteractive.org/topics/measeval/crmref.html

of mastery, or proficiency, on a subject, like mathematics standards. Reporting scores in terms of mastery or proficiency requires some sort of standard-setting process using the assessment to determine just what score(s) represents mastery. Describing these standard-setting processes is beyond the scope of this Learning Point; however, standards are needed to build a test used for providing criterion-referenced scores.

A criterion-referenced score presents information about the level of achievement that a student has demonstrated by their test score. Because the score is compared to the content, it is independent of the people who took the test. Any number of students (or no students at all) might “meet standards” on the test. A single test score, on its own, gives no information about how that student scored compared to others who took the same test.

Norm-referenced scores

In some instances, test results might be more usefully interpreted with respect to other people who took the same test rather than to the content of the test. This need can be met by norm-referenced scores. Rather than providing information about the skills and knowledge that students directly have, norm-referenced scores provide information as to where a particular score ranks compared to scores of other test takers. This gives us answers to questions like “How did this student do compared to other students?”

Norm-referenced test scores are most often reported as a percentile rank. A percentile rank indicates the percentage of test takers who had a score the same as or lower than the observed score. For example, if in 2018 you

scored a 28 on the ACT Reading test, ACT norms tables tell us that your score was equal to or higher than 82% of the people who took the test. Note that a percentile rank is not the same as a percent correct or percentage. Percentile rank is a norm-referenced score and a percentage is a criterion-referenced score.

It is important to know about the group of test takers—the “norm group”—that a norm-referenced score is compared to. In some instances, the norm group may consist of all the people who take a test during an administration. Large-scale assessments will often develop a norm group that is representative of the population of interest. For example, all eighth-graders or perhaps just eighth-graders in urban settings. Scores on subsequent administrations are compared to this group’s scores, even though they did not complete the test at the same time. Characteristics such as recency and representativeness of such a norm group are provided in a test’s technical manual.

It is important to know the characteristics of the norm group for interpretation of the test score. Some large-scale, norm-referenced tests will have different norm groups for different sub-populations. One such example is a test that has two norm groups: one that reflects the county as a whole and another that is representative of an urban population. A given raw score might be associated with different percentiles when compared to these two populations.

Another common norm-referenced score used in educational testing is the grade-equivalent score—possibly the most misunderstood score used in educational testing. Grade-equivalent scores are presented as grade and

month. For example, a grade-equivalent score of 9.2 is interpreted as ninth grade, second month (typically November). A score of 3.0 would be grade 3, September (no months completed in the third grade).

The common misinterpretation of the grade-equivalent score revolves around thinking a score indicates that a student should be placed in another grade, or is doing work at another grade level. For example, if a seventh-grader takes a math test and earns a grade-equivalent score of 9.2, this does not mean that student should be moved to the ninth grade. It also does not mean that the student is doing ninth-grade work. This is because the test the seventh-grade student took was on seventh-grade content.

So how do we interpret a grade-equivalent score? Our seventh-grader did well on this test, since the grade-equivalent score is higher than her actual grade. In fact, our seventh-grade student did as well on the seventh-grade content as we would expect a typical ninth-grade student to do in the month of November. The important thing to remember is that the test is based on seventh-grade content. It doesn’t speak to what a student does or doesn’t know on content taught in other grades.

Large-scale, norm-reference assessments need to be re-normed periodically. This is to reflect changes in achievement in the target population over time.

Summary

Norm- and criterion-referencing provide two different frames of reference with which to interpret test scores. Norm-referenced scores give information about how a student did compared to other students taking the test, but isn’t specific about the student’s strengths or weaknesses in terms of content. Criterion-referenced assessments give more explicit information about the levels of achievement on the content, but don’t communicate how that student did compared to others. Both types of score reports can be useful in different educational situations.

The Michigan Assessment Consortium’s Assessment Learning Network (ALN) is a professional learning community consisting of members from MI’s professional education organizations; the goal of the ALN is to increase the assessment literacy of all of Michigan’s professional educators.