The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation

Criterion-Referenced Interpretation

Contributors: Kyle Nickodem & Michael C. Rodriguez
Edited by: Bruce B. Frey
Book Title: The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation
Chapter Title: "Criterion-Referenced Interpretation"
Pub. Date: 2018
Access Date: February 26, 2018
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks,
Print ISBN: 9781506326153
Online ISBN: 9781506326139
DOI: http://dx.doi.org/10.4135/9781506326139.n166
Print pages: 426-428

Criterion-referenced interpretation is the interpretation of a test score as a measure of the knowledge, skills, and abilities an individual or group can demonstrate from a clearly defined content or behavior domain. It is often defined as a contrast to norm-referenced interpretation, where an individual's score only has meaning when it is compared to other individuals' scores. Criterion-referenced interpretations are independent of information based on how the average person performs. This entry further describes criterion-referenced interpretation and its uses, then discusses the design and validation of tests that foster criterion-referenced interpretation. The entry concludes with a look at common misconceptions about criterion-referenced interpretation.

Criterion-reference interpreted scores have been used for a variety of decisions, such as monitoring student achievement, evaluating efficacy of instructional programs, granting licensure and certification, planning individual and group instruction, and identifying possible learning disabilities. Tests that are designed to foster criterion-referenced interpretation of scores include Advanced Placement assessments, driver's license exams, and the Programme for International Student Assessment.

A criterion-referenced interpretation assumes an underlying continuum of content knowledge and behaviors that ranges from none to all encompassing. When the breadth and depth of knowledge and behaviors that comprise the content domain—the criterion—is clearly and completely specified, and a test is constructed with a representative sample of items from the content domain, it is understood that there is a correspondence between an individual's performance on the test and their ability level on the underlying continuum. Thus, if a test is constructed to foster a criterion-referenced interpretation, the inference can be made that an individual who scores 75% on the test knows and is able to demonstrate individual knowledge of 75% of the content domain.

First outlined by Robert Glaser in his 1963 symposium address to the American Educational Research Association, criterion-referenced interpretation gained popularity in the United States in the 1970s, as the development of theories of measurement and test design refined the distinctions between criterion-referenced and norm-referenced interpretations and their uses. Although it is possible under certain conditions to interpret scores from a single test in reference to both a criterion domain and a norming group, doing so rarely leads to satisfactory interpretations because different score interpretations require different test designs. It is important to note that the nature of test score interpretation (criterion- or norm referenced) is a characteristic of the interpretation as enabled by test design, not the test itself. There is a tendency in the measurement and assessment literature to refer to anything not explicitly norm referenced as criterion referenced. Here, the description of criterion-referenced interpretation is consistent with the original intent.

Design and Validation

As with all well-developed tests, in tests designed to foster criterion-referenced interpretation, the purpose, content domain, test specifications, and item specifications are defined. A key component of the content domain that supports criterion-referenced interpretation is that it covers a relatively narrow set of cognitive skills (although this is not a technical requirement), so that the resulting test sufficiently measures performance within the domain. This requires the test developer to define the boundaries of skills relevant to the content domain as well as the types and formats of problems and scoring rules that delineate membership of appropriate items and tasks. This recognizes natural variability in item difficulty as a function

of conceptual difficulty of items and tasks, the complexity of relevant contexts, and recognition of the natural progress of skill levels in the well-defined domain. The result is a pool of carefully constructed items deeply measuring performance to support criterion-referenced interpretations.

Ideally, once the criterion is well defined, items and tasks are generated that cover the entire expanse of the content domain. From this pool of items, a representative sample is drawn to construct the test. The representative sample of items allows for the correspondence between performance on the test and ability on the underlying knowledge continuum to be established. In practice, however, areas of the domain that are more easily measured tend to be overrepresented, even when they are more peripheral content.

Assuming that items are of high quality, ensuring that the items chosen are a representative sample of the content domain is, theoretically, the only concern regarding item selection for fostering criterion-referenced interpretation. Unlike norm-referenced interpretations, criterion-referenced interpretation does not depend on the variability of scores between test takers. Thus, items that are extremely difficult or extremely easy can be included if they address a fundamental skill or knowledge expected of test takers.

Lack of score variability also means that tests designed to support criterion-referenced interpretation are likely to produce low item-total correlations as measures of item discrimination and result in low internal consistency reliability in a classical test theory sense (thus such estimates are inappropriate for scores intended for criterion-referenced interpretation). Other estimates of score consistency are more appropriate, including decision or classification consistency.

The length of the test is dictated by the scope of the content domain and whether score interpretation is for individuals or groups. The broader the content domain, the longer the test will likely need to be in order for the sample of items to adequately cover the domain. Additionally, individual-level score interpretation requires longer tests because each test taker must respond to items that are representative of the entire content domain. However, group- or program-level score interpretations can be supported with fewer items because the content domain only needs to be appropriately represented when scores are aggregated. This means it is possible for each test taker to only respond to items that cover a portion of the domain as long as the entire domain is covered when aggregated to the group- or program level.

A variety of objective and subjective scoring methods can be used to support criterion-referenced interpretation. Selecting the appropriate scoring method largely depends on the nature of the content domain and the target audience. Although a typical scoring method is to calculate the number or percentage of items answered or tasks performed correctly, this method is not the most meaningful for all criterion domains. For instance, the speed of completing the task, such as running a mile or calculating single-digit multiplication, might be of greater importance, especially when the task itself is relatively easy to complete for the intended population. In other contexts, the precision of performance is of greater interest, as when transcribing an interview or using a rubric to score the quality of a test taker's essay. Many standardized tests employ more sophisticated scoring methods using item response theory or Bayesian estimation along with additional scaling considerations to generate final scores. Regardless of the scoring method employed, the theoretical rational and the procedures used to produce the scores need to be well documented in order to support criterion-referenced interpretation.

Although cut scores or performance standards are not required for criterion-referenced

interpretation, they are often set in order to aid decisions based on the criterion-referenced interpretation of scores. Performance standards or cut scores categorize test takers into two or more performance or mastery levels. For instance, when score interpretations are used for granting certification, a cut score might be set at 85% correct, whereby test takers who answered 85% or more of the items correctly are granted certification. Although rationale and evidence must be provided to justify the use of cut scores, testing standards dictate circumstances under which cut scores can be established and defended, where to set the cut score is a policy decision based on judgment, often supported with empirical information.

Tests designed to support criterion-referenced interpretation are considered quota free, meaning that the number of test takers expected to score above or below the cut score should have absolutely no bearing on where the cut score is set. Instead, just as scores are interpreted in reference to what students are expected to know and do in a clearly defined criterion domain, cut scores should be set with explicit references to the criterion domain, not the relative performance of a reference group.

The primary validity evidence for interpreting scores from a test in reference to a criterion is a carefully and completely defined criterion domain. The criterion is the content knowledge and performance tasks an individual or group from a specified population is expected to know and be able to do under specified circumstances. This involves specifying whether certain skills or knowledge are of greater importance to the domain, whereas others might be more peripheral. Common procedures for defining the criterion domain include gathering judgments from experts in the domain, mutual consensus from a variety of people associated with the domain, and analysis of research and published works in the domain.

Although tests designed for criterion-referenced interpretation often are used to assess what individuals know and can do at the end of an instructional period, the criterion domain can be defined for any point in the instructional process where it might be useful to measure test takers' current achievement. Each component of the test, including purposes, score interpretations, and uses, is subject to validation, where relevant and appropriate evidence is gathered in its defense.

Common Misconceptions

Tests that are specifically constructed to support criterion-referenced interpretation are commonly referred to as criterion-referenced tests; however, this attribution is misleading. Scores from a single test can be interpreted for multiple purposes. For instance, a score could be interpreted both as a measure of what an individual knows and can do (criterion-referenced interpretation) and as a measure of how individual abilities compare relative to other test takers (norm-referenced interpretation). Although some interpretations might be more appropriate than others based on the design of the test, criterion-reference is an attribute of the interpretation of scores and not the test itself.

Another common misconception regards the multiple definitions of the term *criterion*. With the prevalence of tests that utilize cut scores to categorize test takers into performance or mastery levels, many individuals mistakenly refer to the cut score, performance standard, or mastery level as the criterion (e.g., the criterion passing score). However, the criterion refers to the domain of knowledge and behaviors expected from a defined population under specified circumstances. In an attempt to alleviate possible confusion, the term *domain-referenced interpretation* is sometimes used in place of criterion-referenced interpretation (actually, this has been suggested by measurement specialists numerous times but has not been widely

adopted).

A third misconception is treating objectives-referenced interpretations or standards-based assessment interpretations as necessarily criterion referenced. Objectives-based and standards-based score interpretations share many of the measurement and score reporting characteristics as criterion- referenced interpretations in that results offer insight into the behaviors and abilities individuals and groups can currently demonstrate. Many take this similarity to mean that tests designed for objectives-referenced and standards-based score interpretation reveal test takers' knowledge and abilities for specific content domains when, in reality, the scope of the typical standards-based test is far too broad, where many content standards are lightly sampled.

Unlike criterion-referenced interpretation, objectives-referenced and standards-based interpretations do not require as carefully a defined content domain nor items to be a random or representative sample of the domain. Instead, objectives and standards are defined, which themselves are only a subset of the content domain that is expected to be taught. Thus, the inference drawn from the score is no longer what individuals or groups know and can do from the content domain, but what they know and can do from what they were expected to have been taught, in very general terms, because no specific objective or standard is well defined or measured. For school-level accountability, this might suffice as a general indicator; but for individual-level inferences about knowledge, skills, and abilities, this is insufficient.

Kyle NickodemMichael C. Rodriguez
http://dx.doi.org/10.4135/9781506326139.n166
10.4135/9781506326139.n166

Further Readings

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. American Psychologist, 18, 519–521.
Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10(3), 159–170.
Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48(1), 1–47.
Popham, W. J. (Ed.). (1971). Criterion-referenced measurement: An introduction. Englewood Cliffs, NJ: Educational Technology.
Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6(1), 1–9.