

CRITERION-REFERENCED TESTING

Authored by
mohammad looti

November 11, 2025

RECOMMENDED CITATION

mohammad looti (2025). *CRITERION-REFERENCED TESTING*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=68583>

CRITERION-REFERENCED TESTING

Primary Disciplinary Field(s): Education, Psychometrics, Assessment

1. Core Definition

Criterion-referenced testing (CRT) is a method of assessment based upon comparing an individual's performance against a fixed, predetermined standard or criterion, rather than against the performance of a peer group. This fundamental distinction means that a score obtained through CRT indicates the examinee's degree of competence or mastery in a specific domain of knowledge or set of skills, irrespective of how other test-takers performed. The criterion represents an absolute threshold of acceptable performance, often linked directly to learning objectives or professional competency benchmarks.

The core philosophy driving CRT is the principle of measuring mastery. Unlike traditional grading methods that may rely on curved scales or relative class ranking, CRT ensures that every individual's rating is gauged upon the exact same standard, thereby fostering an environment where achievement is defined by internalized ability rather than external competition. This method is often cited as the most fair basis for comparisons when the goal is to determine if a student or professional possesses the necessary foundational skills to move to the next level of instruction or practice safely and effectively.

The criterion itself is carefully defined before the assessment is administered, often taking the form of specific educational standards, behavioral objectives, or legal requirements. For instance, in a medical licensing exam, the criterion might be the ability to correctly diagnose a specific percentage of clinical cases, ensuring that all licensed practitioners meet a minimum standard of public safety and competence. The resultant score is thus a diagnostic statement regarding the examinee's status relative to this specified standard, frequently leading to classification decisions such as "Mastered" or "Not Mastered," or placement into discrete performance levels like "Basic," "Proficient," or "Advanced."

2. Historical Context and Origins

The conceptual foundation for criterion-referenced testing solidified in the mid-20th century, emerging primarily from the need for more systematic and instructionally relevant measurement practices following decades of reliance on broad norm-referenced assessments. The shift coincided with the rise of instructional design principles, particularly the emphasis on defining clear, measurable behavioral objectives, championed by educators like Ralph Tyler and Robert Mager. These educators advocated for assessment tools that could directly confirm whether specified instructional goals had been achieved.

The term "criterion-referenced testing" was formally introduced into the psychometric lexicon by Robert Glaser in 1963. Glaser, recognizing the limitations of norm-referenced measures for judging the effectiveness of instruction, argued that tests should provide information about what an individual can actually do and what they know, rather than merely how they compare to others. His work provided the theoretical framework necessary to develop tests designed explicitly to assess student performance against an absolute scale of achievement, linking assessment results directly back to specific educational content.

The widespread adoption and practical application of CRT were significantly bolstered by the development of Mastery Learning models, primarily popularized by Benjamin Bloom in the late 1960s and early 1970s. Mastery learning necessitated assessment tools capable of determining whether a learner had fully acquired prerequisite knowledge before advancing. CRT provided the ideal measurement instrument for this purpose, offering clear, unambiguous feedback about competence and ensuring that assessments were fundamentally aligned with the instructional objectives they were intended to measure, marking a major departure from assessments primarily designed for selection or ranking.

3. Distinguishing CRT from Norm-Referenced Testing

The primary distinction between criterion-referenced testing and norm-referenced testing (NRT) lies in the purpose and interpretation of the scores. While CRT aims to describe performance relative to a fixed standard (the criterion), NRT aims to describe performance relative to a defined peer group (the norm). This difference dictates the construction, administration, and utility of the assessment, influencing everything from item selection to score reporting.

In NRT, test items are often selected specifically to maximize variance in scores, ensuring a wide distribution that facilitates ranking and differentiation among individuals; items that everyone answers correctly or incorrectly are often excluded because they do not contribute to rank ordering. Conversely, in CRT, item selection is driven solely by content validity--items must accurately sample the domain of knowledge specified by the criterion. If all examinees correctly answer an item that measures a critical, required skill, this is considered a successful outcome for CRT, as it indicates widespread mastery of that specific standard.

Interpretation of results further highlights the difference. An NRT score (e.g., a percentile rank) tells an employer that an applicant scored better than 85% of their peers, but it reveals nothing about the actual knowledge or skills the applicant possesses. Conversely, a CRT score (e.g., a "Proficient" designation) tells the employer that the applicant has successfully met 90% of the required industry standards. This absolute standard is why CRT scores remain stable regardless of the composition of the test-taking group; a high-achieving cohort does not make it harder to pass, nor does a low-achieving cohort make it easier, a central element noted in its reputation for

fairness.

4. Purposes and Applications

Criterion-referenced testing is indispensable in situations where specific, measurable competencies must be verified, particularly when public accountability or safety is involved. One of the most common applications is in high-stakes educational testing, where states or countries mandate that students achieve a defined level of proficiency in core subjects (e.g., reading, mathematics) before graduation or promotion. These assessments ensure that the educational system is upholding minimum standards for all learners.

Beyond the K-12 environment, CRT forms the backbone of virtually all professional licensing and certification programs. Whether testing prospective airline pilots, doctors, or plumbers, the governing bodies must guarantee that every certified professional meets a non-negotiable standard of competence. The ability to fail or pass is not dependent on how the rest of the cohort performs, but purely on whether the individual has achieved the established criterion necessary for safe and effective practice.

Furthermore, CRT is highly valued for its diagnostic utility in clinical and instructional settings. Because the test items are meticulously mapped to specific learning objectives, a CRT score profile can pinpoint precisely which standards an individual has met and which standards require further instruction. This specificity is crucial for tailoring individualized educational plans (IEPs) or designing targeted professional development programs, as it moves assessment feedback away from generalized ability statements and toward actionable instructional data.

5. Measurement and Interpretation

The technical rigor of criterion-referenced testing hinges upon the process of **standard setting**--the formal procedure used to establish the cut score or threshold separating acceptable performance from unacceptable performance. This process is complex, often relying on structured judgment methods involving panels of subject matter experts (SMEs). Popular methods include the Angoff method, the Bookmark method, and the Contrasting Groups method, each designed to systematically elicit expert opinion on the minimum performance required to achieve mastery.

Interpretation of CRT results generally falls into two categories: dichotomous and polytomous. Dichotomous interpretation results in a simple classification, such as "Pass/Fail" or "Competent/Not Competent." This binary outcome is typical for certification exams where meeting the single cut score is the only required outcome. Polytomous interpretation involves establishing multiple cut scores that divide the performance scale into several distinct levels, such as "Below Basic," "Basic," "Proficient," and "Advanced." These levels provide a more nuanced understanding of the examinee's relationship to the criterion domain.

A key technical consideration in CRT is ensuring the test adequately samples the entire domain of knowledge defined by the criterion. This requirement necessitates meticulous blueprinting and test construction to maintain high content validity, ensuring that the test items genuinely represent the standards they are intended to measure. The precision of the measurement, particularly near the cut score, is critical, as misclassifying an individual (e.g., classifying a competent person as non-competent, or vice versa) carries significant professional or educational consequences.

6. Advantages of Criterion-Referenced Testing

One of the most significant advantages of CRT, directly supporting the assertion from the source material that it is the "most fair basis for comparisons," is its inherent equity. Since success is not competitive, the performance of an individual is judged solely on their own merits against a fixed, public standard. This promotes transparency in the assessment process and mitigates the possibility of systemic unfairness that can arise when grading systems are forced to conform to a pre-defined distribution curve, regardless of overall class achievement.

Furthermore, CRT possesses superior **instructional utility** compared to NRT. The results derived from criterion-referenced assessments are actionable because they are diagnostic, directly linking specific student deficiencies to specific, failed instructional objectives. Educators can use this precise feedback to adjust their pedagogy, remediate specific learning gaps, and confirm the effectiveness of curricular changes, thereby closing the assessment-instructional loop efficiently and effectively.

The use of CRT strengthens external accountability systems. When educational institutions report scores based on defined criteria (e.g., 85% of students achieved Proficiency in Algebra I), these scores communicate clearly defined expectations to stakeholders, including parents, legislators, and employers. This clarity contrasts sharply with percentile ranks, which require context about the norm group to be meaningful. The explicit nature of the criterion allows for stronger alignment between teaching, testing, and desired outcomes, driving instructional practice toward achieving necessary competencies.

7. Criticisms and Limitations

Despite its advantages, criterion-referenced testing is subject to several methodological and practical criticisms. The most pressing challenge is the inherent subjectivity and complexity involved in **standard setting**. Determining where the cut score should definitively lie--the point separating mastery from non-mastery--is often a highly political and judgmental process, not a purely mathematical one. Critics argue that small, arbitrary changes in the placement of the cut score can lead to massive differences in passing rates, which raises concerns about the reliability and validity of the consequential decisions based on those scores.

Another significant limitation relates to the high cost and labor intensity of test development. Producing high-quality CRT items requires exhaustive efforts to ensure that the test content perfectly samples the defined domain and that the items function reliably as indicators of the criterion. This detailed mapping and piloting process is often more expensive than developing NRT assessments, which focus primarily on maximizing score dispersion rather than content specificity.

Finally, while excellent for diagnosing specific mastery, CRT offers limited ability for large-scale comparative evaluation unless the standards themselves are perfectly consistent across different contexts and time periods. If two states utilize CRT but set different proficiency standards for the same subject, their results cannot be directly compared to determine which state's students are performing "better" overall. This lack of inherent comparative data can be a political drawback when stakeholders require relative performance metrics for funding decisions or national ranking purposes.

8. Further Reading

[Criterion-referenced test \(Wikipedia\)](#)

[Glaser, R. \(1963\). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18\(9\), 519-521.](#)

[American Psychological Association \(APA\). Standards for Educational and Psychological Testing.](#)