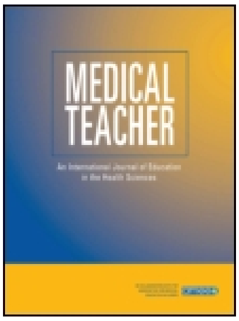

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

# What Is... Normative versus Criterion-referenced Assessment

Jeffrey M. Turnbull

# *What Is . . . Normative versus Criterion-referenced Assessment*

**JEFFREY M. TURNBULL,** *Department of Medicine, University Hospital and University of Western Ontario*

SUMMARY *Any reform of the current medical curriculum requires a reappraisal of the principles and practices of the evaluation system. The norm-referenced test which at this time is the primary method of evaluation was initially designed to rank order students for the purpose of selection. Difficulties arise when it is used, as it currently is for the assessment of competence. The norm-referenced test is often insensitive to instruction and, while it provides information regarding the relative strengths and weaknesses of students in comparison to their peers, it does not provide an estimate of the absolute level of performance achieved. In addition to promoting competition among students, the norm-referenced test is less suitable for programme evaluation.*

*As it is the principal responsibility of a medical school to produce competent physicians and not to rank order them, it is more reasonable to compare student achievement to an external standard of performance or criterion. Criterion-referenced testing, then, is more suitable for the assessment of competence and, within this setting, percentage competency scores can be utilised when there is a need for the rank ordering of student achievement for the purposes of selection.*

*It is recommended that criterion-referenced testing be the primary method of evaluation, as it best meets the objectives of the medical school by emphasising the achievement of clearly established external standards and, thereby, ensuring a high quality of performance.*

## Introduction

For three decades, educational theorists have argued the relative merits of the norm-referenced test (NRT) versus the criterion-referenced test (CRT) (Ebel, 1971; Block, 1971(a); Swezey, 1981; Popham, 1978(a); Airasian & Madaus, 1974). While initially a topic of intense educational research and debate among academicians such as Gagné, Bloom, Popham, and Ebel, this long simmering controversy then spread into our schools and universities. More recently, it has reappeared in the forum of medical education where it has a special relevance. For medical educators this is not a discussion that focuses solely on student evaluation. It challenges us to consider how we teach, why we teach, and for whom do we teach.

There is a trend in medical education which encourages independent learning, de-emphasises memorisation of factual material and, at the same time, expands our clinical and problem-solving skills. Equally, there is a greater sense of accountability to our students and the public. If our curriculum is to evolve in this direction, there must be a corresponding alteration in the way we assess our students as the evaluation process plays a pivotal role in curriculum development. As we reassess many of the principles of medical education, we must also reconsider the existing evaluation process. It is because of this, that a preview of the controversy between the NRT and the CRT is not only relevant but essential.

Before proceeding further, it is important to outline the constituent parts of this debate, providing definitions and an understanding of the fundamental principles involved. All tests are relative in that they are referenced to some standard. A raw score is meaningless unless it can be related to some other standard of performance such as a percentage correct or a percentile standing in a class. The distinction between norm-referenced testing and criterion-referenced testing centres around the standard for comparison and the subsequent ramifications that result from this.

The standard of comparison in a NRT is a cohort whose scores are distributed along a normal Gaussian curve. The NRT, then, is designed to ascertain the student's status in relation to the performance of this group who have completed the test. Here, the basic tenet is that one's peers are used to set the standards for the assessment of *comparative* ability and *relative* attainment. To illustrate this, consider a norm-referenced biochemistry test in the preclinical years. This test would be designed so that class performance is distributed along a normal curve with the mean adjusted in order to allow a fixed percentage failure. Subsequently, this would permit one to rank order students in comparison to their peers.

A CRT is one where a student's abilities are compared to a pre-set standard of performance. The CRT ascertains an individual's status with respect to this external criterion or performance standard. Here, the basic tenet is that the student's absolute performance is assessed which is not relative to his or her peers but to some pre-set criterion determined by the faculty. An example of a CRT is the same biochemistry examination; however, before the test is initiated, a series of clearly-defined criteria, felt to be necessary for the successful completion of the course, are established. The achievement of these performance criteria or a fixed percentage of them determines the student's success or failure. The CRT is less suitable for the rank ordering of students but permits a statement of the absolute level of performance achieved.

With greater appreciation for the principles behind the NRT and the CRT, a more detailed assessment of their relative merits and limitations as well as their potential role within medical education can be considered.

## Norm-referenced Testing

Norm-referenced testing has enjoyed great popularity over the last 50 years, much of which is deserved. It flourished at a time when aptitude testing was prominent (the Stanford Binet IQ score being a classic NRT) and has continued to dominate educational assessment. The NRT is prevalent throughout medical education and, the

Canadian and American licencing exams are norm-referenced. It has well established standards for validity and reliability and it captures the prevailing attitude of many educators that students are not judged by absolute measures but by comparison to others. This usually results in a fixed percentage of students who fail or achieve an excellent standing. In addition, the NRT is suitable for the comparison and ranking of students for the purposes of selection such as internship selection.

There are, however, definite limitations to the NRT which can be classified as those where evaluation impacts upon instruction and those specific to the area of evaluation alone. From an instructional point of view, there are several serious limitations to norm-referenced testing. Firstly, the NRT depends upon variance. Student differences must be amplified if reliable compartive scoring is to take place. Topics of importance which are emphasised, taught well, and evaluated in a clear manner result in a high degree of success on the examination. As these questions do not discriminate, they are disregarded. Often the discriminating norm-referenced test item is one that is ambiguous or vague, and tests areas not taught. The selection of these test items results in a test that is insensitive to instruction.

The second instructional concern with norm-referenced testing is that it fails to provide a clear understanding of just what the student can or cannot do. In the above biochemistry test, a percentage mark on a NRT does not tell the student, faculty, nor the public what performance standards have been met and, in fact, the percentile or percentage mark can vary depending upon who constitutes the norm-referenced group or cohort. For the assessment of absolute standards of achievement and student feedback related to these standards, the NRT is unsuitable, as strengths and weaknesses are lost in the overall performance of the class.

Finally, the examination process is a powerful incentive to learn. The NRT misdirects this motivation as students compete against their peers unnecessarily for passing grades rather than working together in an effort to achieve essential competencies.

Traditionally, the domain of the NRT has been the assessment of factual knowledge. Medical curricula are evolving away from factual recall towards independent learning and problem-based learning. The NRT has been used infrequently for the assessment of these traits and its utility in this area is unknown.

Norm-referenced testing is also limited in providing useful information for the purpose of programme evaluation. There is a growing demand on the part of the public and the government for accountability in medical education. In the area of programme evaluation, the NRT is of questionable value and has been criticised by Swezey (1981) and Popham (1978b) for providing meaningless and misleading data. There is no area more important than the education and training of a physician where clear standards or criteria for achievement are essential for all to examine. The NRT cannot discern to what degree an educational programme has met these standards whereas criterion-referenced testing can.

## Criterion-referenced Testing

Recognising the norm-referenced test's limitations, Glasser (1963) formalised the concept of criterion-referenced testing. Its early acceptance, however, was in part due to the emerging emphasis placed upon clearly stated behavioural objectives. The development of a CRT entails, firstly, a statement of behavioural objectives and then a systematic generation of test items designed to unambiguously ascertain to what degree these objectives have been met. Standards of performance are set using minimal levels of competence before the test is applied. Following the test there is no attempt to alter the percentage pass or fail as the standards of performance are based upon uncompromising pre-set objectives.

Criterion-referenced testing addresses many of the shortcomings of the NRT in that there is no longer a need for variation, as students are not compared to their peers. Consequently, the examination process becomes more responsive to what is actually taught. Clearly specified objectives and achievement standards allow us to know exactly what the student is capable of and where his or her deficiencies lie with regards to remedial action. In addition to clarifying teaching, there is data presented by Duchastel (1973) and Block (1971b) to suggest that the pre-specification of objectives and achievement standards also enhance learning.

Criterion-referenced testing compares student performance to a preset criterion and not to the class average. Consequently, it challenges the existing practice of adjusting marks to guarantee a fixed percentage of students who will fail or achieve honours. Medical students have already been rank ordered and selected. It is the medical educator's task to motivate students to a high level of uniform competence and, as such, the percentage failures should depend upon the achievement of mandatory criteria and not the relative class standing.

Unfortunately, criterion-referenced testing is not without its difficulties and limitations. One perceived difficulty is that the CRT does not encourage excellence but merely guarantees that the majority of the class will achieve the minimal acceptable standard. High achievement standards in criterion-referenced testing have been shown by Block (1972) to maximise group excellence. Individual excellence can also be enhanced as students who have demonstrated an acceptable level of competence are free to pursue additional interests in greater depth.

Another perceived difficulty of the CRT is that an absolute standard of pass or fail does not give adequate constructive feedback for the purpose of self-evaluation. A CRT is the summation of a series of test items felt to adequately assess knowledge, skills and, possibly, attitudes. If, for example, out of one hundred stated objectives seventy-six are met and the minimal standard is set at eighty-five then that student has not met the criteria for the minimal acceptable performance; however, his or her percent competency score (76%) relative to the minimal standard (85%) is known and specific items that were answered incorrectly can be drawn to the student's attention.

Finally, there are concerns that if the CRT is endorsed then difficulties will arise with the selection of medical graduates for internship purposes. This is not a stated objective of most faculties of medicine. However, this could be achieved by the comparison of percent competency scores. Equally, the CRT would permit a statement of the absolute abilities of our graduates making the selection process not only easier but more valid.

There are, however, several discrete limitations of criterion-referenced testing as summarised by Ebel (1971). Criterion-referenced tests are expensive, time-consuming, and difficult to produce with an accurate index of reliability. While they have been used with success in the assessment of basic intellectual skills, they are untried in the assessment of more complex, cognitive tasks such as problem-solving. Finally, the setting of clear instructional objectives, as we all know, is difficult. Equally complex, however, is the development of appropriate minimal levels of competency. The setting of this minimal standard is beyond the scope of this article; however, it is the subject of several publications (Van der Linden, 1982; Hambleton *et al.*, 1978; Millman, 1973).

## Validity and Reliability

The measure of any psychometric test depends upon its validity and reliability. It is only reasonable that we now compare the NRT with the CRT in these terms. Validity or the degree to which a test is measuring what it is intended to measure has several components of which content validity is the most important. Content validity is the detailed analysis of test items and the degree to which they match the intent of the assessment process. As there is often a mis-match between what is taught and what is examined in norm-referenced testing, there is a lack of content validity. Criterion-referenced testing, on the other hand, systematically matches test items to instructional objectives and has higher content validity.

Reliability or the consistency with which a test measures what it is intended to measure is often assessed differently between norm-referenced testing and criterion-referenced testing. While standards of stability with time or equivalence with another test can equally be applied to the NRT or the CRT, methods of assessing internal consistency, however, frequently cannot. The NRT is the prototype upon which the parametric analysis of internal consistency (as measured by the Kuder-Richardson formula or the more generalisable Cronbach alpha coefficient) has been developed. Unfortunately, the results of the CRT are infrequently distributed in a Gaussian fashion making parametric statistical analysis inaccurate (Lehmann & Mehrens, 1984; Shavelson *et al.*, 1972; Subkoviak, 1984). Non-parametric analysis is still in the state of development and, as a result, stability and equivalence may be the best standards to test the reliability of an individual CRT. These measures are cumbersome and often the CRT is not accompanied by a detailed estimate of its reliabilty.

## Conclusions

The controversy surrounding norm-referenced testing and criterion-referenced testing focuses attention upon more fundamental issues than the comparison of two simple psychometric tests. These two methods of assessment are not inherently right or wrong. They are, however, often unknowingly applied in inappropriate circumstances and interpreted incorrectly. The fundamental question that must be asked is what is the purpose of evaluation in medical education today and in the future? As the principal objective of medical education is to produce a competent physician, then, unquestionably, the *raison d'être* for the evaluation process is to assess the standard of competency and not the rank order of students. To this end, criterion-referenced testing is necessary and must become the principal method of evaluation within medical education. Norm-referenced testing is currently used because of its ease, relatively low cost, our familiarity with it, and our lack of understanding of its usefulness. It is severely limited when used for purposes other than the ranking of students for selection. It should be noted, however, that the CRT is not a panacea nor is it value free, and the limitations of criterion-referenced testing such as cost and the development of non-parametric statistical estimates of reliability remain.

What is required is the endorsement of criterion-referenced testing in medical education. This will lead to the development of comprehensive objectives and their corresponding performance criteria. Criterion-referenced testing must then be used extensively in the assessment of all necessary competencies and subsequent critical review will allow the prediction of its utility in areas such as the assessment of problem-solving and essential attitudes. In addition, further development is necessary of a better statistical estimate of the realiability of an individual CRT and the setting of mastery scores. The setting of minimally acceptable competency scores or mastery scores is recognised as a powerful educational incentive; however, in setting mastery scores we must be cognisant of the relative risks of failing a competent student versus passing one who is incompetent. With these developments and the subsequent implementation of criterion-referenced testing, this system of evaluation promises to be more useful and representative.

Educational theorists have tended to underestimate the potential role of the evaluation system to alter the shape and direction of the curriculum. It is apparent that any alteration of the evaluation process will have major implications throughout the curriculum. Consequently, the debate between norm-referenced testing and criterion-referenced testing is paramount as we reconsider changing our approach to medical education. The principal objective of medical education is best met by criterion-referenced testing as it emphasises achievement of clearly established external standards ensuring a high quality of performance within our graduates.

*Correspondence:* Dr J. M. Turnbull, Department of Medicine, University Hospital, Box 5337, Station A, London, Ontario N5A 5A5, Canada.

## REFERENCES

AIRASIAN, P.W. & MADAUS, G.F. (1974) Criterion-referenced testing in the classroom, in: TYLER, R.W. & WOLF, R.M. (Eds) *Crucial Issues in Testing*, pp. 73–88 (Berkley, CA, McCutcheon).

BLOCK , J.H. (1971a) Criterion-referenced measurements: potential, *School Review*, 79, pp. 289–297.

BLOCK, J.H. (1971b) *Mastery Learning: theory and practice* (New York, Holt, Rinehart & Winston).

BLOCK, J.H. (1972) Student learning and the setting of mastery performance standards, *Educational Horizons*, 50, pp. 183–190.

DUCHASTEL, P. C. (1973) The effects of behavioural objectives on learning: a review of empirical studies, *Review of Educational Research*, 43, pp. 53–67.

EBEL, R.L. (1971) Criterion-referenced measurements: limitations, *School Review*, 79, pp. 282–287.

GLASER, R. (1963) Instructional technology and the measurement of learning outcomes, *American Psychologist*, 18, pp. 519–521.

HAMBLETON, R.K. *et al.* (1978) Criterion-referenced testing and measurement: a review of technical issues and developments, *Review of Educational Research*, 48, pp. 1–47.

LEHMANN, I.J. & MEHRENS, W.A. (1984) Measurement and evaulation, in: *Education and Psychology*, pp. 275–310 (New York, Holt, Rinehart & Winston).

MILLMAN, J. (1973) Passing scores and test lengths for domain-referenced measures, *Review of Educational Research*, 43, pp. 205–216.

POPHAM, W.J. & HUSEK, T.R. (1971) Implications of criterion-referenced measurement, in: POPHAM, W.J. (Ed.) *Criterion-referenced measurement: an introduction*, pp. 17–37 (New Jersey, Educational Technology).

POPHAM, J. (1978a) *Criterion-referenced measurement* (Englewood Cliffs, NJ, Prentice Hall).

POPHAM, J. (1978b) The case for criterion-referenced measurements, *Educational Research*, pp. 6–10.

SHAVELSON, R.J. *et al.* (1972) Criterion-referenced testing: comments on reliability, *Journal of Educational Measurement*, 9, pp. 133–137.

SUBKOVIAK, M.J. (1984) *Estimating the Reliability of Mastery-non-mastery Classifications*, ed. by BERK, R.A. (Baltimore, MD, Johns Hopkins University Press).

SWEZEY, R.W. (1981) *Individual Performance Assessment: an approach to criterion-referenced test development* (Virginia, Reston).

VAN DER LINDEN, W.J. (1982) Criterion-referenced measurement: its main applications, problems, and findings, *Evaluation in Education*, 5, pp. 97–117.