# SETTING STANDARDS FOR CRITERION-REFERENCED MEASUREMENT

*By :*

**Dr. MAHMOUD M. GHANDOUR**

QATAR UNIVERSITY

Improving human resources is the responsibility of the educational system in many societies. The outputs of the educational systems are, in many ways, the inputs of other systems found in societies. Life today in many societies is complicated and heavily dependent upon technology. In my opinion, the world societies today can be divided into two categories the countries which produce technology and the countries which consume technology.

Since technology is considered to be a product of education, educational systems must be continuously active. A review of literature indicates that one area in which education is active today is the field of educational measurement. Society is demanding that educational systems be accountable; educational measurement is one way of trying to meet that demand.

Many new theories and concepts have been advanced in the field of measurement, with new techniques being adopted to support these new theories and concepts. One of the main emphasis of educational measurement today is criterion-referenced measurement, where an individual's score is compared to a well established standard of behavior. Other labels given to this testing movement are minimum competency testing, proficiency testing, mastery learning, domain-referenced testing, and objective-referenced testing.

The issue of criterion-referenced testing has occupied the concern of specialists in the united states since the early 1970's when a shift began from norm-referenced testing, where scores derived from test scores of a

norm group are used to make comparative judgments or statements about an individual. Burton (1978, p. 263) states that the major objective behind this movement is "to transfer responsibility for some important educational decisions from individual teachers to a more uniform, more scientific technology." Levin (1978) writes that this movemnt seems to be a natural extension of schooling in an industrial society because schools, like private enterprises that produce goods and services, are expected to produce or achieve certain results. However, if this belief is held, then it must be possible to ascertain what the outputs should be and to assess both institutional and student performance. Levin concludes that it is not surprising that these premises are rarely questioned by educators.

Robert Glass (Glass, 1978) first used the term "criterion referenced test" in a paper written in 1972 on assessing human performancs. In his 1963 classic essay, "Instructional Technology and Measurement of Learning Outcomes," he differentiated between norm-referenced measurement and criterion-referenced measurement. Glass used the concept of norm-referenced measurement to describe achievement tests that discern an examinee's relative standing and the concept of criterion-referenced measurement to describe tests that identify an examinee's absolute mastery or non-mastery of specific behaviors.

Since the term "criterion-referenced measurement" was first coined by Glaser, more than fifty definitions or descriptions have appeared in research (Berk, 1980). Comparisons of these definitions suggest a general agreement that the test is used to reference an examinee's score to a well-defined domain of behaviors. Popham's definition (Popham, 1978, p. 73) captures the essence of most of the other descriptions : "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavior domain.

The major problem facing the criterion-referenced test construction is the setting of standards. Hambleton (1978, p. 279) defines a standard or a cut-off score as "a point on a test score scale that is used to sort examinees into two categories that reflect different levels of proficiency relative to a particular objective measured by a test."

Jaeger (1976) writes that standard setting is a judgmental act. Nothing can replace the final judgmental act of deciding which performa-

nces are acceptable and which are unacceptable. Because standard setting procedures depend on different information and varying degrees of judgment, proficiency standards differ with the method used.

This paper will address the topic of standard setting and the different methods or techniques used to identify cut-off scores.

Glass (1978) identifies six classes of techinques used to determine the criterion score on a criterion-referenced test : performance of others, counting backwards from 100%, bootstrapping on other criterion scores, judging minimal competence, decision-theoretic approaches, and operations research methods. A brief description of each of these methods follows.

**Performance of others as a Criterion :** The parameters of existing populations of examinees can be used to establish criterion levels. The median test score earned by persons of a certain type can be used as the standard cut-off score. Since this technique is, in fact, pure norm-referencing, several criterion-referenced test theorists consider it to be an inappropriate method (Hambleton et al., 1978).

**Counting Backwards from 100% :** An objective is written and test items are written to correspond to it. A performance of 100% is desired; however, it is recognized that perfection is impossible and concessions need to be made. These concessions are arbitrary with some allowing a 5% shortfall and others allowing 20% or more.

**Judging Minimal Competence :** Experts study a test and then declare what score a "Minimally competent" person should score. Glass feels this approach fails because the concept of minimal competence has no foundation in psychology and because judges cannot agree on what constitutes minimal competence.

**Bootstrapping on other Criterion Scores :** Criterion scores are set by articulation with a passing score (success or mastery) on some other examination or external judgment. Glass concludes that external tests or judgments permit no sensible nonarbitrary demarcation of scores into categories such as skilled, unskilled or knowledgable, ignorant.

**Decision-Theoretic Approaches :** Persons are divided into two groups

according to some external criterion of interest suck as employed versus not employed. These same persons are administered a criterion-referenced test and a criterion score established where they can be classified as passing or failing. Four categories of passing or failing the criterion-referenced test and external criterion is possible. Using the decision-theoretic technique, the cut-off score on the criterion-referenced test would be allowed to vary in order to vary the proportions of persons in each category. This allows one to minimize or maximize the consequences of setting the criterion score at a certain level. The weighting of false positives and false negatives is arbitrary.

**Operations Research Method** : This technique is based on maximizing a valued commodity by locating an optimum point on a mathematical curve or graph.

Poggio, Glasnapp and Eros (1981) compared four frequently used standard setting methods : Angoff, Ebel, Nedelsky, and Contrasting Groups. The results of the study indicated that the Ebel method produced the highest standard, then Angoff, then Contrasting Groups, and the lowest standard was produced by the Nedelsky method. A brief explanation of each method follows :

**Angoff Method** : In this method, judges estimate the difficulty level of an item using as a reference a hypothetical group of minimally competent individuals. This standard represents the estimate mean total score for the hypothetical group of minimally competent individuals.

**Ebel Method** : In this method, judges rate each item according to level of difficulty (3 levels) and level of relevance (4 levels). After each item is rated, the judges indicate what percentage of items must be answered correctly within each of the 12 cells to be judged minimally competent. The items are then assigned to each cell to get a standard.

**Nedelsky Method** : This approach can only be used with multiple choice tests. Judges indicate the detractors that a minimally competent student should be able to eliminate as incorrect for each item. The score represents the item's difficulty level and the standard represents the mean total test score expected from the reference group. This method also allows the

users to determine what percentage of minimally competent students should fall above and below the standard.

**Contrasting Group Method** : Students who have scores on a test are classified into two groups-competent or not competent-according to the content being measured. A standard can be derived by using the group membership and actual test scores and by using the statistical likelihood ratio procedure which minimizes the probability of misclassification of students into groups.

Koffler (1980) also compared standards derived from the Nedelsky prcedure and the Contrasting Groups procedure. The results showed that the cut-off scores from these two procedures were different. Koffler concluded that no one procedure should be relied on to set cut-off scores, but that a number of procedures should be used.

Burton (1978) reviews three widely accepted methods for setting standards : theories, expert judgments, and practical necessity.

**Theories** : The concept of criterion-referenced testing was originally closely related to learning hierarchies. Today much is being done with non-hirarchical learning theories, but these theories are not broad enough to be considered general decision making tools yet. Burton rejects theoretical procedures because learning hierarchies have not been established and other theories appear to be too limited.

**Expert Judgments** : When theory is lacking, standards can be based on the experience of experts. Classroom teachers are the most appropriate experts as they have access to most of the relevant information. Burton rejects expert consensus beyond the classroom level, because the required level of information is not available.

**Practical Necessity** : In the early 1970's, the concept of performance standards began to include minimal competence. The idea developed that if one could identify the skills needed in everyday life, then its practical value as a skill is justification as a standard. Burton rejects practical necessity techniques because there are many causes of reallife successes. There is no single skill so necessary that survival depends upon it.

Burton concludes from her review of the aforementioned methods for setting performance standards, that there is no practical performance stan-

dards technology today and that the potential of such technology is limited and not a promising vehicle for social decision-making.

Levin (1978) discusses three different methods used to construct educational performance standards : Pedagogical approach, pragmatic approach, and scientific approach.

**Pedagogical Approach** : This approach is probably the dominant one used up to this time (Glass, 1978). Educators determine domains which they consider important for students to achieve and then, with the help of testing experts, construct test items to measure the domains. Judgments tend to be arbitrary and results reflect the formal curriculum of the school.

**Pragmatic Approach** : This method attempts to establish minimal tasks which adults should be able to perform in our society.

**Scientific Methodology** : This approach entails a systematic search for adult requirements and then selecting standards on the basis of how well they predict mastery of these competencies.

Levin concludes that none of these three approaches allows one to construct a defensible set of performance standards for certifying student competencies except in the most arbitrary sense.

The most serious technical problem which might face criterion-referenced measurement is the reduced variance because everyone wants to achieve the criterion. Consequently, there is no further application of other statistical techniques such as tests or analysis of variance. Reliability cannot be measured because the major function of this process is to bring every person to the criterion level. This also means that the validity cannot be estimated.

I agree that the educational process needs certain decisions made at specific stages. However, these decisions need to be objective ones, as science deals with objectivity rather than subjectivity. In setting the standards for criterion-referenced measurement, one can assume there is subjectivity, especially when we deal with opinions of judges who set their standards haphazardly or in an arbitrary manner.

The educational process is a multivariate phenomena which includes several variables. In order to use criterion-referenced measurement, equal oportunity must be secured for every student. This cannot happen; even if we could control the curriculm, we could not control the teachers and the teaching methods used.

When the criterion is strictly applied, what happens to the students on the border. Are we going to let these students pass or will we retain these students at the same level? Can the child retake the test immediately and improve his/her score so as to pass? There appears to be many unanswered questions? Even if these questions are answered with a yes or no, on what basis are we justifying our answers?

My position on this issue is that there is no absolute usage of either norm-referenced measurement or criterion-referenced measurement. Each is suitable for a specific situation and certain needs. We can use norm-referenced measurement when we want to compare students according to their relative positions; however, this should not be the only criteria used. We can use criterion-referenced measurement when we monitor the achievement of students regarding certain specific objectives. Again, this should not be the only criteria used. Thus major problem concerning criterion-referenced measurement is the arbitariness of setting the standards. The decision rules need to be stated on how criterions are set; also, school should experiment with using more than one standard setting method to see how the cut-off scores vary. Educational personnel need to try to make standard setting as objective or as informative as they can. I agree with Burton that the decisions from criterion-referenced measurement should not affect areas beyond the classroom. Criterion-referenced measurements are great for diagnostic purposes but appear to be too unobjective for summative purposes.

## BIBLIOGRAPHY


Berk, Ronald A. ed. **Criterion-Referenced Measurement : The State of the Art** The John Hopkins University press : Baltimore and London, 1980, p. 5.

Burton, Nancy W. "Societal Standards", **Journal of Educational Measurement.** 15 (4), Winter 1978, p. 237-261

Glass, Gene W. "Standard and Criteria" **Journal of Educational Measurement.** 15 (4), Winter 1978, p. 237-261.

Hambleton, Ronald A. "On the Use of Cut-off Scores with Criterion-Referenced Tests in Instructional Settings", **Journal of Educational Measurement.** 14 (4), Winter 1978, p. 277-290.

Hambleton, R.K.; Swaminathan, H.; Algina, J., & Coulson, D.B. "Criterion-Referenced Testing and Measurement : A Review of Technical Issues and Developments. **Review of Educational Research.**

Jaeger, R. M. "Measurement Consequences of Selected Standard-Setting Models. " **Florida Journal of Educational Research.** 1976, 18, p. 22-27.

Koffler, Stephen L. "A Comparison of Approaches for Setting Proficiency Standards. **"Journal of Educational Measurement.** 17 (3), Fall 1980, p. 167-178.

Levin, Henry M. "Educational Performance Standards : Image or Substance". **Journal of Educational Measurement.** 15 (4), Winter 1978, p. 309-319.

Pggio, John P.; Glasnapp, Douglas R.; & Eross, Down S. "An Emperical Investigation of the Angoff, Ebel, and Nedelsky Standard Setting Methods". (Presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, April, 1981).

Popham, W. James. **Criterion-Referenced Measurement.** Prentice-Hall Inc. : Inglewood Cliffs, New Jersey, 1978.