February 2003

Standards and Criteria Redux

Gene V Glass
College of Education
Arizona State University

[Note: I published the paper "Standards and criteria" in the *Journal of Educational Measurement* in 1978 (Vol. 15, 237-261). I am now revisiting it because the message is more urgent now than it was then. This paper is essentially a reprinting of the 1978 paper, to which I plan to add a prologue.]

A common expression of wishful thinking is to base a grand scheme on a fundamental, unsolved problem. Politicians outline energy policy under the assumption that physicists will soon be able to control the intense heat generated by nuclear fusion. Planners chart the future course of cancer research with faith that basic discoveries will be made at an expenditure of $2 billion plus or minus. Those who think on exalted levels are prone to underrate the complexity of what seem lesser problems. Utilitarianism in ethics is an example. "The greatest good for the greatest number" is not only logically inconsistent —since one can't maximize two functions simultaneously—but as a social policy, it falls at the final hurdle: there exists no social calculus by which one can compute the amount of good eventuating from a social policy.

Contemporary educational movements present a similar situation: accountability, mastery learning, assessment, competency-based education, minimal competence graduation requirements. A literature search under any one of these categories brings a deluge of reports, speeches, and position papers. The movements have spawned laws, jobs, conferences, and distinguished commissions. And, much of the language and thinking rests at bottom on a common notion: that a minimal acceptable level of performance on a task can be specified. Whether it goes by the name "mastery," "competence," or "proficiency," it is the same fundamental notion. A judge (technician, professional, and the like) inspects an exercise or task or test and somehow determines that the score C_x represents mastery, minimal competence, proficiency, etc.. A recent incident in New England could be a bellwether for school districts across the country:

> By a vote of 6 to 2, the board of education in Stamford, Conn., has adopted a resolution requiring applicants for teaching jobs to "demonstrate mastery of written and spoken English as a pre requisite to being hired." The resolution also stipulated that teachers now employed in the Stamford schools would be tested in English and those found "deficient in communication" would receive remedial instruction.

I have read the writings of those who claim the ability to make the determination of mastery or competence in statistical or psychological ways. They can't. At least, they cannot determine "criterion levels" or standards other than arbitrarily. The consequences of the arbitrary decisions are so varied that it is necessary either to reduce the arbitrariness, and hence the unpredictability of the consequences of applying the standards, or to abandon the search for criterion levels altogether in favor of ways of using test data that are less arbitrary and, hence, safer.

This monograph has grown out of a series of discussions and a six-month period of reading and reflecting on the literature which were initiated by Fritz Mosher's suggestions to the National Assessment of Educational Progress (NAEP) to examine the "standards" question. Conversations with Mosher himself and the staff of NAEP have been most influential. The Analysis Advisory Committee of NAEP, under Fred Mosteller's chairmanship, proved a rigorous testing ground for many of the ideas.

In the following pages, I shall (a) examine the ordinary usage of the words "standards" and "criteria" in the measurement literature; (b) trace the evolution of the notion of performance standards in the criterion-referenced testing movement; (c) analyze and critique six methods of setting performance standards on criterion-referenced tests; and (d) reflect briefly on the political forces which have become focused on the standards issue.

"Standards" In Common Parlance

Setting standards or mastery levels is frequently written about as though it is a well-established and routine phase of instructional development. In conversations with measurement specialists and instructional development experts over the past few years, I have been literally dumbfounded by the nonchalance with which they handle the standards problem. One will report that he always sets a standard of two-thirds of the items correct for mastery because he's a sort of "liberal guy." Another expert will report that he holds learners to 70% mastery, and a third advances his 90% standard with an air of tough-mindedness and respect for excellence. None of them bothers with such apparently extraneous considerations as how the test items are to be composed and whether they will be abstruse or obvious. In one of the sacred writings of the instructional objectives movement, Robert F. Mager (1962) identified standard setting as an integral part of stating an objective properly:

> If we can specify at least the minimum acceptable performance for each objective, we will have a performance standard against which to test our instructional programs; we will have a means for determining whether our programs are successful in achieving our instructional intent. What we must try to do, then, is indicate in our statement of objectives what the acceptable performance will be, by adding words that describe the criterion of success. (p. 44)

Mager went on to illustrate what he meant by a behavioral objective and its associate standard:

- The student must be able to correctly solve at least seven simple linear equations within a Period of thirty minutes.

- Given a human skeleton, the student must be able to correctly identify by labeling at least 40 of the. . . bones; there will be no penalty for guessing.
- The student must be able to spell correctly at least 80 percent of the words called out to him during an examination period. (p. 44)

This language of performance standards is pseudoquantification, a meaningless application of numbers to a question not prepared for quantitative analysis. A teacher, or psychologist, or linguist simply cannot set meaningful standards of performance for activities as imprecisely defined as "spelling correctly words called out during an examination period. And, little headway is made toward a solution to the problem by specifying greater detail about how the questions, tasks, or exercises will be constructed. Can a more meaningful performance standard be stated for an objective as molecular as "the pupil will be able to discriminate the grapheme combination 'vowel + r' spelled 'ir' from other graphemes"? Can it be asserted confidently about this narrow objective that a pupil should be able to make 9 out of 10 correct discriminations? In point of fact, this objective appears on the Stanford Reading Test where it is assessed by two different items:

> **a) Mark the word "firm" (Read by proctor)**

firm	form	farm

> **b) Mark the word "girl" (Read by proctor)**

goal	girl	grill

The percentages of second-grade pupils in the norm population answering items a) and b) correctly were 56% and 88%, respectively. Any performance standards (e.g., "8 out of 10 correct") for a group of items like item **a** would be quite inappropriate for a group of items like item **b**, since they are so different in difficulty. Results from a grade seven assessment by the Department of Education in New Jersey illustrate the same point. Pupils averaged 86% on vertical addition, but only 46% on horizontal addition. The vagaries of teaching and measurement are so poorly understood that the *a priori* statement of performance standards is foolhardy.

Benjamin S. Bloom (1968), whose name has become closely associated with the notion of "mastery learning," has written of instructional psychology in ways that depend fundamentally on notions of performance standards:

> Most students (perhaps over 90 percent) can master what we have to teach them. (p. 1)
>
> There is little question that the schools now do provide successful learning experiences for some students—perhaps as high as one third of the students. If the schools are to provide successful and satisfying learning experiences

> for at least 90 percent of the students, major changes must take place in the attitudes of students, teachers, and administrators... (p.2)
>
> Thus, we are expressing the view that, given sufficient time (and appropriate types of help), 95 percent of students...can learn a subject up to a high level of mastery. We are convinced that the grade of *A* as an index of mastery of a subject can, under appropriate conditions, be achieved by up to 95 percent of the students in the class. (p. 4)

Popham (1973), writing on instructional objectives for teachers in training, reaffirmed the centrality of performance standards:

> There is, however, another dimension to objective writing, a dimension that further aids the teacher in planning and evaluating his instruction. It involves establishing performance standards, that is, specifying prior to instruction the minimal levels of pupil achievement. (p. 3)

The notion of performance standards is repeatedly illustrated in Popham's teachers' manual:

> In a math class, the student will be able to solve ten of fifteen perimeter problems. (p. 3)
>
> The student will be able to identify correctly, through chemical analysis procedures, at *least five* unknown substances. (p. 6)

Wiersma and Jurs (1976), in outlining the instructional evaluation component of Individually Guided Education (the University of Wisconsin R & D Center instructional plan), gave the following description of criterion-referenced testing:

> When an individual's performance score is interpreted with reference to an established criterion and without reference to the level of the performance of a group, we have a criterion referenced interpretation. The criterion is usually established prior to any actual measurement being done. The criterion or criteria are usually stated in the instructional objectives or in supplements to the stated objectives. For example, a list of objectives may have an accompanying statement indicating that when students score 90 percent correct on the related test, they should be considered as having attained the objectives. (p. 14)

In detailing the role of testing in assessment programs, Ralph W. Tyler (1973) illustrated a performance standard for determining mastery:

> For example, in primary reading, the children who enter without having learned to distinguish letters and sound might be tested by the end of the year on letter recognition, association of letters with sounds, and word-recognition of one hundred most common words. For each of these specified "things to be learned," the child would be presented with a large

> enough sample of examples to furnish reliable evidence that he could recognize the letters of the alphabet, he could associate the appropriate sounds with each letter, alone and in words, and he could recognize the one hundred most common words. A child has demonstrated mastery of specified knowledge, ability, or skill when he performs correctly 85 percent of the time. (Some small allowance, like 15 percent, is needed for lapses common to all people.) (p. 105)

The staff of the National Assessment of Educational Progress have grappled with the performance standards problem for years to almost no one's satisfaction. Though they have never adopted an official position on the matter, they did cooperate with the National Council for the Social Studies in an effort to apply performance standards to the assessment results in citizenship and social studies (Fair, 1975). A fully representative panel of nine judges (3 minorities, 5 women, 3 under the age of 30) was formed. Each judge was shown an assessment item and then asked, "Realistically what level of performance nationally for the age level being considered would satisfy you for this exercise? (1) less than 20% correct, (2) 20-40%, (3) 41-60%, (4) 61-80%, or (5) more than 80%?" The panel rendered over 5,000 judgments in a three-day sitting, and it has been reported that "...panel members agreed more often than not, but at times spread their responses across all the available categories" (Fair, 1975, p. 45). About half of the exercises were given a "satisfactory performance level" of "more than 80%." About 35% of the exercises would satisfy the panel if between 60% and 80% of the examinees answered correctly. The desired performance levels were generally above the actual rates of correct response. What is to be made of the gap? Ought it to be read as evidence of the deficiency of the educational system; or is it testament to the panel's aspirations, American hustle and the indomitable human spirit ("Man's reach Should exceed his grasp, etc.")?

The reader can justifiably ask, "What manner of discourse is being engaged in by these experts?" How is one to regard such statements as "the student must be able to correctly solve at least seven simple linear equations in thirty minutes" or "90 percent of all students can master what we have to teach them." If such statements are to be challenged, should they be challenged as claims emanating from psychology, statistics, or philosophy? Do they maintain something about learning or something about measurement? Are they disconfirmable empirical claims or are they merely educational rhetoric spoken more for effect than for substance?

The Evolution of "Criterion-referenced Testing"

An historical digression can contribute much to clarifying the evolution of the contemporary notion of a "criterion-referenced test." The first known use of the term "criterion-referenced test" was made by Robert Glaser in a chapter on assessing human performance, which was coauthored by David Klaus and published in a book edited by Robert Gagne in 1962. This initial treatment of the topic antedated by a year the widely read and better known publication by Glaser, "Instructional Technology and the Measurement of Learning Outcomes" in the *American Psychologist*, 1963.

Glaser (1963) sought to emphasize the importance of making scores informative about behavior rather than merely about relative performance on poorly specified and vaguely known dimensions assumed to lie behind a test score:

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on this continuum as indicated by the behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum. The term "criterion," when used in this way, does not necessarily refer to final end-of-course behavior. Criterion levels can be established at any point in instruction where it is necessary to obtain information as to the adequacy of an individual's performance. The point is that the specific behaviors implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels. It is in this sense that measures of proficiency can be criterion-referenced.
>
> Along such a continuum of attainment, a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others. (pp. 519-520)

There were in Glaser's early writings a few intimations that criterion-referenced tests could be used in establishing cut-off scores between competence and incompetence or that such distinctions as pass-fail and mastery-nonmastery make psychological sense. Rather, as the quotation above reveals, there is assumed to be "... a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance" and the "*... degree of competence attained by a particular student*" [emphasis added] is what is assessed. Competence is conceived of as being a continuum characteristic. There are, at most, ambiguous suggestions that a single point exists at which competence becomes incompetence. Only once in his early paper did Glaser (1963) lapse into the rhetoric of cut-off scores:

> We need to behaviorally specify minimum levels of performance that describe the least amount of end-of-course competence the student is expected to attain, or that he needs in order to go on to the next course in a sequence. (p. 520)

At nearly the same time that Glaser was developing his thoughts about criterion-referenced measurement, Mager (1962) published what was soon to be his widely read and highly influential exposition on behavioral objectives, *Preparing Instructional Objectives*. The passage in Mager's (1962) text most pertinent to tracing the development of contemporary ideas of criterion-referenced testing was cited above and is repeated here:

> If we can specify at least the minimum acceptable performance for each objective, we will have a performance standard against which to test our instructional programs; we will have a means for determining whether our programs are successful in achieving our instructional intent. (p. 44)

Thus, Mager added the idea of the performance standard to the long standing notion of the behavioral objective.

The writings of both Glaser and Mager were influential in the development of testing and evaluation during the mid-1960s. Among the persons significantly influenced by both was W. James Popham. Indeed, Popham seems to have played a primary role in amalgamating the language of Glaser and Mager. In 1969, Popham and Husek wrote one of the most often cited papers on criterion-referenced testing. They wrote of "criterion-referenced measurement," and they used Mager's term "performance standard":

> Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., performance standard. (p. 2)

Glaser's use of the word "criterion" with its colloquial meaning of "standard," the simultaneous publication of Mager's rather simple notions of performance standards, and Popham's mixing of Glaser and Mager in the same pot combined to create the impression that the "criterion" in criterion-referenced testing was not the behavioral scale articulated to a test and elaborating the meaning of the scores, but rather that the "criterion" was the cut-off score—the division between pass & fail, mastery & nonmastery, competence & incompetence. This interpretation of the word "criterion" is evident in the informal conversation of both educators and measurement specialists. This meaning is intended when people speak, as they do now habitually, of "setting the criterion on a criterion-referenced test or test item." Furthermore, it is clear that statisticians and psychometricians who have addressed themselves to the mathematical analysis of criterion-referenced tests have had this meaning of "criterion" in mind. They axiomatize the criterion referenced testing problem as follows: "Consider a score C_x on a test such that those persons with true scores above C_x are said to 'pass' the test."

When Glaser and Nitko (1971) sought to clarify the meaning of "criterion-referencing" some eight years after Glaser's original papers, the notion of a performance standard crept in at the end of the definition:

> A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified

> performance standards.... The performance standards are usually specified by defining some domain of tasks that the student should perform. Representative samples of tasks from this domain are organized into a test. Measurements are taken and are used to make a statement about the performance of each individual relative to that domain. (p. 653)

The concept of a performance standard was absent from Harris and Stewart's (1971) definition of a criterion-referenced test: "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed. (p. 1) Iven's (1970) definition similarly avoided any suggestion of a performance standard and non-comparative evaluation: A criterion-referenced test is one "comprised of items keyed to a set of behavioral objectives" (p. 2). Lindvall and Nitko (1975) listed four defining characteristics of a criterion-referenced test, none of which suggests a performance standard or cut-off score:

> ...there are four characteristics inherent in criterion-referenced tests:
>
> 1. The classes of behaviors that define different achievement levels are specified as clearly as possible before the test is constructed.
> 2. Each behavior class is defined by a set of test situations (that is, test items or test tasks) in which the behaviors and all their important nuances can be displayed.
> 3. Given that the classes of behavior have been specified and that the test situations have been defined, a representative sampling plan is designed and used to select the test tasks that will appear on any form of the test.
> 4. The obtained score must be capable of being referenced objectively and meaningfully to the individual's performance characteristics in these classes of behavior. (p. 76)

In the activity of the early 1970s, it was largely forgotten that the first principles of criterion-referenced testing were uncertain and tentative. The belief became widely accepted that criterion-referenced tests carry with them a performance standard or cut-off score indicating mastery. By 1976, the "cut-off score" interpretation of criterion-referenced testing had advanced so far that at an AERA symposium entitled Criterion-Referenced Testing, four of the five papers were essentially psychometric treatments of the cut-off score problem (AERA 1976 Annual Meeting Program, p. 187, Session 27.03).

Glaser's thinking after his seminal 1963 paper has evolved in a direction of fuller appreciation of the complex and variegated fabric of behavior and testing. Glaser's choice Of the term "criterion" was quite sensibly suggested by the use of the term in classic psychometrics. There the word "criterion" denoted a measurement scale used in validating a test or psychometric scale. It is generally a scale formed by the observation or recording of behavior which the psychometric instrument is to predict. For example, the psychometric test might be a paper-and-pencil vocational interest inventory, and the criterion, a scale of persons' actual occupational choices. Or, the test could be performance on a form board, and the criterion, an evaluation of employees' speed and accuracy in operating a cash register.

It was in this classic psychometric sense that Glaser (personal communication, 1976) intended the term "criterion" in criterion-referenced testing to be understood. He envisioned tests closely articulated to the relevant behaviors which traditional psychometrics embodied in the criterion scale but seldom in the test itself.

The evolution of the meaning of "criterion" in criterion-referenced tests is, in fact, a case study in confusion and corruption of meaning. We find that a careful reading of Glaser's thoughts on the nature and use of criterion-referenced testing is compelling, and they contain little of Mager's suggestion that performance standards will be created *ex nihilo* and be used to decide mastery or nonmastery. The coincidence in time of Glaser and Mager's work, and Popham's enthusiastic purveying of both positions have created the contemporary confusion of the two. Furthermore, the indiscriminate mixing of Glaser and Mager's thinking has lent the force of Glaser's cogent observations about testing to Mager's less defensible recommendations about "performance standards."

Jackson (1970) probably best described Glaser's current conception of criterion-referenced testing when he wrote, "... the term 'criterion-referenced' will be used here to apply only to a test designed and constructed in a manner that defines explicit rules linking patterns of test performance to behavioral referents" (p. 3)." It is the mathematicians and other simplifiers who prematurely translated a tentative notion—one that must wait for the development of a more sophisticated instructional and learning psychology—into the idea of "cutoff scores" and "mastery levels." If ever there was a psychological-educational concept ill-prepared for mathematical treatment, it is the idea of criterion-referencing.

(Several persons who read earlier drafts of this monograph feared that criticism of methods of establishing standards or cut-off points might be carelessly read as criticism of associated notions that are logically separate, most notably "domain referenced testing." I was persuaded that a warning was needed. But where to place it is a problem; one can't predict where someone might draw an unwarranted association. The warning will have to fit here whether or not it seems the proper place. The objections raised against criterion-referenced tests up to this point and beyond concern the notion of a cut-off score, standard, or criterion level. They do not apply to notions of domain referenced testing nor to any other of a number of eminently sensible suggestions for writing tests.)

Methods of Determining the Cut-score

Questions of the intended meaning of "criterion-referenced test" aside, we must deal at length with the work that has been spawned by the corrupted meaning of the word "criterion," i.e., the sense of criterion as a standard, mastery level, cut-off score, or pass-fail mark. The word "criterion" is for now taken as synonymous with "standard" or "cut-off" and not in the sense of a scale of behavior loosely linked or articulated to a test scale.

We have identified six classes of technique for determining the criterion score on a criterion-referenced test. Five of these methods are very nearly the same as the methods of establishing cut-offs that Millman (1973) identified.

1. Performance of Others;
2. "Counting Backwards from 100%";
3. Bootstrapping on Other Criterion Scores;
4. Judging Minimal Competence;
5. Decision-Theoretic Approaches;
6. "Operations Research" Methods.

Performance of Others as a Criterion

Some criterion levels are established by reference to parameters of existing populations of examinees. Hence, the criterion or mastery level on a test may be established as the median test score earned by persons of a certain type. There are a few prominent examples of this method of criterion setting.

The California High School Proficiency Examination was created as an instrument to determine whether students above age 16 are to be certified (not "graduated") and released from high school. The implementation of this examination created the problem of setting a passing score. It was determined that the 50th percentile of graduating seniors would constitute the criterion score. Thus, the criterion was determined normatively and not by direct reference to the behaviors exhibited on the test (only in so far as the behaviors are reflected in the 50th percentile).

In Arizona, a senior-year examination was instituted for potential graduates. A proficiency level was established as 9.0 grade-equivalent units on a standardized reading achievement test. But since 9.0 grade-equivalent units on the achievement test is a scale score defined as the median score earned by ninth-graders in September, what might appear to some to be a behaviorally informative score is, in fact, merely normatively determined. (This example, cited in 1977, is *not* the Arizona AIMS test of the 1990s. That test, which failed as many as 85% of the students who took the high-school exit exam, used a different method of establishing a cut-score. It remains one of the most inept attempts to handle the cut-score problem ever recorded. 2/14/2003)

These examples reveal that using the performance of others in these ways to establish a criterion score is, in fact, pure norm-referencing; and, thus, as a means of setting the criterion, this must surely be a mild embarrassment to criterion-referenced testing proponents who have so often attempted to build their own house by tearing down that of the norm-referenced testers.

"Counting Backwards from 100%"

Many criterion scores appear to have been established in a manner appropriately, though perhaps facetiously, referred to a "counting backwards from 100%." An objective is stated and a test item is written to correspond to it. Since the objective is felt to be important—or else it wouldn't have been stated—its author readily endorses the proposition that everyone should be able to answer the test question based on it, i.e., the "desired performance level" is 100%. But reason and experience prevail and it is quickly recognized that perfection is impossible and concessions must be made for mental infirmity, clerical errors, misinformation, inattention, etc. Just how great a concession is to be made becomes distressingly arbitrary with some allowing a 5% shortfall and others allowing 20% or more. For example, "A child has demonstrated mastery of specified knowledge, ability or skill when he performs correctly 85 percent of the time. (Some small allowance, like 15 percent, is needed for lapses common to all people." (Tyler, 1973, p. 105)

If the criterion is set in terms of percent of test items (e.g., 95% of these items will be answered by each student, then the arbitrariness in counting backwards from 100% can have even more serious consequences. If Expert A sets the criterion at 95% and Expert B sets it at 90%, the difference in the percent of examinees attaining the two different criterion levels can vary greatly (say, from 10% in the former case to 50% in the latter). Where one stops counting (e.g., at 99% or 95% or 80%) manifestly controls the percent deemed to have reached the criterion. But the difference between failing 5% and failing 25% of the pupils may be crucial; and if so, it ought not to be decided by a judgmental process so subject to whim and idiosyncrasy as this one.

Bootstrapping on Other Criterion Scores

In this technique—seldom if ever employed to my knowledge, but quickly suggested by a consideration of the problem—a criterion score on a test is determined by articulating the test with an external designation of "success" or "mastery." For example, one might first identify those candidates for the bar (or for certification as barbers, cosmetologists, actuaries, realtors, dentists, and so on) who successfully achieved certification. This group, then, is a group of "competent" persons, judged so by other means. By studying the distribution of their scores on the test in question, perhaps a criterion score can be established on the test for separating the competent from the incompetent.

There are a least two problems with this technique. First, suppose that an examination was given to prospective realtors and the realtor licensing board established the cut-off score. If the second test on which the criterion-referenced tester wanted to establish the criterion score is less than perfectly correlated with the licensing exam (as it most certainly would be), then any cut-off score on the criterion-referenced test (CRT) will be exceeded by some of the licensed realtors but not exceeded by others.

The positioning of the criterion score on the criterion-referenced test cannot be made so that there is perfect correspondence between those who pass the licensing exam and those who pass the criterion-referenced test. Thus, it becomes arbitrary where on the criterion-referenced test the cut-off is drawn. The arbitrariness can be partially disguised by adopting decision-theory techniques of minimizing or maximizing various cost functions of "false-negatives" and "false-positives," (See Figure 1) but it will never be eliminated. (The decision-theory approach to setting the criterion score is discussed under *Decision-Theoretic Approaches* below.

The second difficulty with setting criterion scores on criterion-referenced tests by articulation with a passing score on some other examination or outside judgment is that in so doing, one, in effect, begs the question of the possibility of setting such a standard in the first place. One might well ask, "How does the licensing agent rationalize his choice?" If the choice can be rationalized, then the methods by which it was derived should be identifiable, and thus they could be applied to the problem-of setting the criterion score on the criterion-referenced tests.

When one inquires into what methods are used to set cutting scores on such instruments as civil-service tests, licensing examinations, etc., one finds that the methods have little to do with psychological-behavioral analysis. Contrary to popular conception, civil service examinations do not have "pass" scores; rather, the candidates are examined, their scores ranked, and one counts down-from the top of the list of examinees until all of

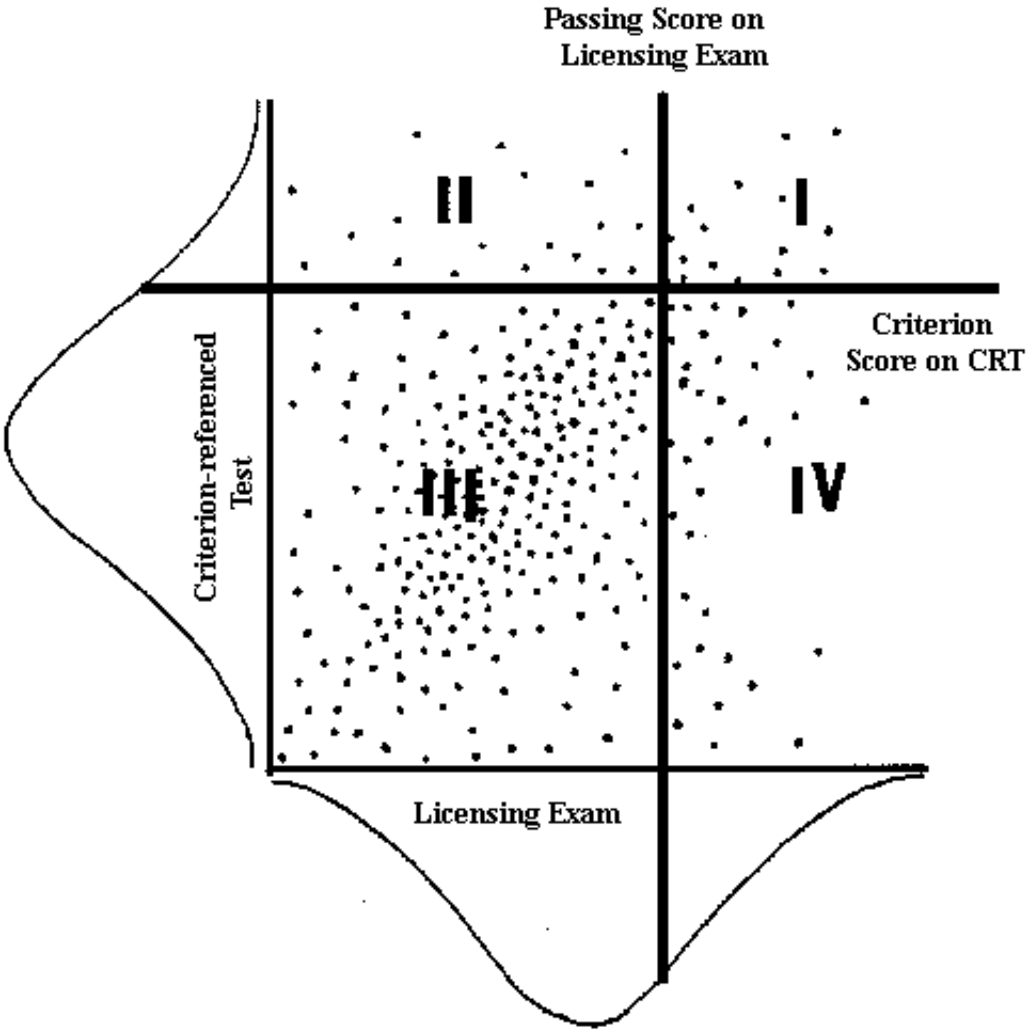


Figure 1. Relationship between a criterion-referenced test and an external examination. (Note: Persons in Quadrant II pass the CRT but fail the licensing exam. Persons in

Quadrant IV pass the licensing exam but fail the CRT.)

the available jobs are filled. Written examinations for licensing automobile drivers have passing scores, usually at around 90% of the questions. Whether the number of errors permitted is 2 or 5 or 10 is completely arbitrary, and there is scant reason to believe that highways would be less safe if the permissible error rate on the test were doubled or tripled. "Passing" scores on licensure exams (for barbers, dentists, physicians, psychologists, etc.) are governed almost exclusively by principles of supply and demand for manpower in the labor market. These cut-off points have virtually nothing to do with defensible judgments of competent vs. incompetent. Thus, it is as if one reached to lift himself by his bootstraps and found none there.

I am *not* maintaining that licensing or personnel selection tests are uncorrelated with valid, important criteria, in the classic psychometric sense. They usually are and by law (Griggs vs. Duke Power Co.) must be. I am maintaining, however, that these tests permit no sensible, non-arbitrary demarcation of scores into two categories described by words and ideas like "competent vs. incompetent," "skilled vs. unskilled," "knowledgeable vs. unknowledgeable."

Judging Minimal Competence

In this approach, experts study a test or an item or an exercise and then declare that a "minimally competent" person should score such-and-so. This has been the direction taken in legislation in Oregon and New Jersey in attempts to control graduation from high school. Two refinements of this technique are due to Nedelsky (1954) and Ebel (1972).

Nedelsky outlined his technique as follows:

> "The proposed technique for arriving at the minimum passing score on an objective test, each item of which has a *single* correct response, is as follows:
>
> *Directions to Instructors*
>
> "Before the test is given, the instructors in the course are given copies of the test, and the following directions:
>
> "In each item of the test, cross out those responses which the lowest D-student should be able to reject as incorrect. To the left of the item, write the reciprocal of the number of the remaining responses. Thus if you cross out one out of five responses, write 1/4.
>
> > Example. (The example should preferably be one of the items of the test in question.)
> >
> > > Light has wave characteristics. Which of the following is the best experimental evidence for this

statement?

- Light can be reflected by a mirror.
- Light forms dark and light bands on passing through a small opening.
- A beam of white light can be broken into its component colors by a prism.
- Light carries energy.
- 1/4 ~~Light operates a photoelectric cell.~~

Preliminary Agreement on Standards

"After the instructors have marked some five or six items following the directions above, it is recommended that they hold a brief conference to compare and discuss the standards they have used. It may also be well that at this time they agree on a tentative value of constant *k* (see section on The Minimum Passing Score). After such a conference the instructors should proceed independently.

Terminology

"In describing the method of computing the score corresponding to the lowest D the following terminology is convenient:

- Responses which the lowest D-student should be able to reject as incorrect, and which therefore should be primarily attractive to F-students, are called *F-responses*. In the example above, response E was the only F-response in the opinion of the instructor who marked the item.
- Students who possess just enough knowledge to reject F-responses and must choose among the remaining responses at random are called *F-D students*, to suggest border-line knowledge between F and D.
- The most probable mean score of the F-D students on a test is called the *F-D guess score* and is denoted by $\mathbf{M}_{\mathbf{FD}}$. As will be shown later, $\mathbf{M}_{\mathbf{FD}}$ is equal to the sum of the reciprocals of the numbers of responses other than F-responses. (In the example above, the reciprocal is 1/4.)
- The most probable value of the standard deviation corresponding to $\mathbf{M}_{\mathbf{FD}}$ is denoted by $\mathbf{Sigma}_{\mathbf{(FD)}}$.

"It should be clear that "F-D students" is a statistical abstraction. The student who can reject the F-responses for every item of a test and yet will choose at random among the rest of the responses probably does not exist; rather, scores equal to $\mathbf{M}_{\mathbf{FD}}$ will be obtained by students whose patterns of responses vary widely.

The Minimum Passing Score

"The score corresponding to the lowest D is set equal to $\mathbf{M_{FD} + k(Sigma_{(FD)})}$, where $\mathbf{M_{FD}}$ is the mean of the $\mathbf{M_{FD}}$ obtained by various instructors, and *k* is a constant whose value is determined by several considerations. The F-D students are characterized not so much by the positive knowledge they possess as by being able to avoid certain misjudgments. Most instructors who have used the F-D guess score technique have felt that this "absence of ignorance" standard is a mild one, and that therefore the minimum passing score should be such as to fail the majority of F-D students. Assigning to *k* the values -1, 0, 1, and 2 will (on the average) fail respectively 16 percent, 50 percent, 84 percent, and 98 percent of the F-D students. An informed final decision on the value of *k* can be reached after the instructors have chosen the F-responses, for at that time they are in a better position to estimate the rigor of the standards they have been using. In keeping within the spirit of absolute standards, however, the value of *k* should be agreed on before the values of $\mathbf{M_{FD}}$ are computed and certainly before the students' scores are known.

"It is the essence of the proposed technique that the standard of achievement is arrived at by a detailed consideration of individual items of the test. Only minor adjustments should be effected by varying the value of *k*. The reason for introducing constant *k*, with the attendant flexibility and ambiguity, is that F-responses in most examinations vary between two extremes; the very wrong, the choice of which indicates gross ignorance, and the moderately wrong, the rejection of which indicates passing knowledge. If a particular test has predominantly the first kind of F-responses, this peculiarity of the test can be corrected for by giving *k* a high value. Similarly, a low value of *k* will correct for the predominance of the second kind of F-responses. It is expected that in the majority of cases a change of not more than +/-.5 in the tentative value of *k* agreed upon during the preliminary conference should introduce the necessary correction. It would be difficult to find a theoretical justification for values of *k* as high as two; for more tests the value *k* = 0 is probably too low. This suggests a rather narrow working range of values, say between .5 and 1.5 with the value *k* = 1 as a good starting point.

"If a part A of a given test consists of $\mathbf{N_A}$ items, each of which has $\mathbf{S_A}$ non F-responses (one of these being the right response), the F-D guess score for each item, i.e., the probability that an F-D student will get the right answer in any one item, is $\mathbf{p_A = 1/S_A}$. The most probable values of the mean and the square of the standard deviation on this part of the test are given by $\mathbf{M_A = p_A(N_A)}$ **and** $\mathbf{(Sigma)^2{}_A = p_A(1 - p_A)N_A}$. $\mathbf{M_{FD}}$ **= Summation-over-A of** $\mathbf{M_{FD,A}}$; and $\mathbf{(Sigma)^2{}_{FD}}$ **= Summation-over-A of** $\mathbf{(Sigma)^2{}_A}$. The value of $\mathbf{M_{FD}}$ must be accurately computed for each test. $\mathbf{Sigma_{(FD)}}$, however, may be given an approximate value. In a test of five-response items, *s* may

> vary from one to five. If these five values are equally frequent, **Sigma$_{(FD)}$** = .41(N). If, on the other hand, the extreme values, *s* = 1 and *s* = 5, are less frequent than the other three values, as seems likely to be true for most tests, .41($N^{.5}$) < **Sigma$_{(FD)}$** < .50($N^{.5}$). Since **k(Sigma$_{(FD)}$)** is usually much smaller than **M$_{FD}$**, approximations are in order. With *k* = 1 and **(Sigma$_{(FD)}$)** = .45($N^{.5}$), the equation, **Minimum Passing Score = M$_{FD}$ + .45($N^{.5}$)**, should work out fairly well in the majority of cases and is therefore recommended as a starting point in experimenting with the proposed technique." (pp. 4-7)

Ebel's (1972) technique is as follows:

> The second weakness of the definition of the passing score as some percentage of the total score is that it still leaves substantial elements of chance in determination of the passing score. The items may be more difficult, or less difficult or less discriminating, than the test constructor intended. Whether an examines passes or fails a specific test may be determined by the questions in the test rather than by his level of professional competence.
>
> The second weakness of this approach can be overcome to some degree by the derivation of the passing percentage from a subjective analysis of the relevance and difficulty of each item in the test. Table 19.7 illustrates four categories of relevance and three categories of difficulty, and gives the expected percentages of passing for items in each category. These expected percentages are what would be expected of a minimally qualified (barely passing) applicant.

Table 19.7. Relevance, Difficulty, and Expected Success on Test Items

	Difficulty Levels		
Relevance Categories	**Easy**	**Medium**	**Hard**
Essential	100%	_	_
Important	90	70%	_
Acceptable	90	60	40%
Questionable	70	50	30

> Suppose, for example, that the number of items in a 100-item test falling in each category when the ratings of five judges are pooled were as shown in the second column of Table 19.8. The sum of these products divided by 500 gives an estimate of the appropriate passing score.

Table 19.8. Passing Score Estimated from Item Characteristics

Item Category	Number of Items*	Expected Success	(Number)X(Succcess)
Essential	94	100	9400
Important			
Easy	106	90	9540
Medium	153	70	10710
Acceptable			
Easy	24	80	1920
Medium	49	60	2940
Hard	52	40	2080
Questionable			
Easy	4	70	280
Medium	11	50	50
Hard	7	30	210
	500		
	37130	=74.26	
	500	or 74%	=passing score

* Actually the number of placements of items in the category by all five of the judges. (Pp. 493-494)

Angoff (1971) presented a technique essentially equivalent to Ebel's but which did not bother with relevance-by-difficulty breakdowns of the items:

> ... ask each judge to state the probability that the "minimally acceptable person" would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or porportions, would then represent the minimally acceptable score. (p. 515)

There are two potential problems: (a) Can judges make such determinations consistently and reliably?; (b) What is the logical-psychological status of the concept of "minimal competence"?

Little empirical research has been reported on the first problem. But a solid, recent study produced startling findings. Andrews and Hecht (1976) carried out an empirical comparison of the Nedelsky and Ebel methods. A group of eight judges was selected from among a committee of individuals who had contributed 180 four-option items to a multiple-choice examination which was nationally administered for certifying professional workers. The judges met on the two separate occasions to set standards once by the Nedelsky method and then by the Ebel method. The study was carefully designed with counterbalancing of order and halves of the test to control for order and memory effects. The findings were astounding. By the Ebel method, the percentage of questions which in the opinion of the judges should have been answered correctly by a "minimally competent" person was 69%. The corresponding percentage determined by the Nedelsky method was 46%. This difference is disconcertingly great. However, the situation is more serious than even a 23-point gap in percentage of items correct would indicate. This percentage difference in number of items correct required to "pass" the certifying examination does not indicate directly the difference in percentages of examinees who would "pass" the test by the Ebel 69% (of items correct) criterion versus the Nedelsky 46% criterion.

We can estimate these two percentages of examinees who "pass" by making a few reasonable assumptions. (The following calculations were not performed by Andrews and Hecht (1976), and they may not vouch for the assumptions on which the calculations are based. Nonetheless, they seem reasonable to me.) Assume that the 180 test items are of average difficulty, i.e., p = .50 for each item; then the mean of the 180-item test would be 90. Furthermore, assume that the range of scores is from a chance score to a perfect score, and that the distribution of total scores is roughly normal. Under these conditions, the standard deviation of the total test scores would equal about one-sixth the range, so that Standard Deviation = (Perfect Score-Chance Score)/6 = (180 - 45)/6 = 135/6 = 22.5.

One can estimate roughly, then, that the total test scores probably have a normal distribution with mean 90 and standard deviation 22.5. This distribution is depicted in Figure 2 where the Ebel and Nedelsky "passing scores" are also indicated.

The figure reveals an enormous discrepancy between the Ebel and Nedelsky standards. Only 7% of the examinees would be certified by the Ebel standard, whereas 63% of the examinees would be certified using the Nedelsky standard. The impression of scientific objectivity created by the rigmarole of grids and guessing corrections quickly evaporates when one sees the staggering discrepancy between the pass rates of the two standard setting methods.

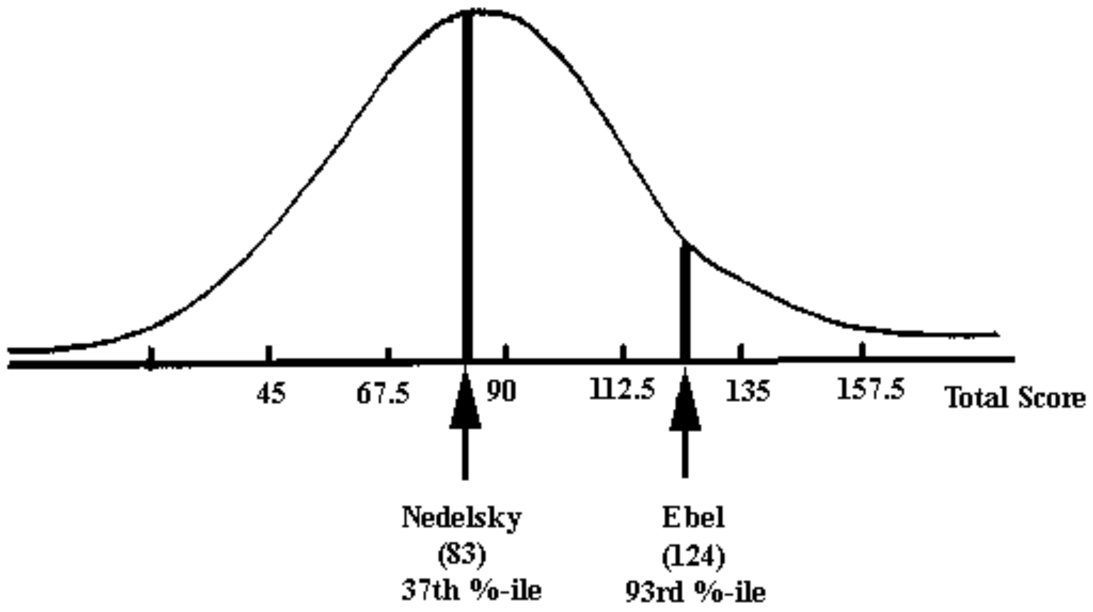


Figure 2. Comparison of the Nedelsky and Ebel cut-off scores.

The logical and psychological status of the concept of minimal competence must be questioned. The history of toxicology presents a case of the fruitless use of an analogous concept, the "minimal lethal dose." The concept was discarded by Trevan (1927) nearly fifty years ago:

> The common use of this expression [minimal lethal dose] in the literature of the subject would logically involve the assumptions that there is a dose, for any given poison, which is only just sufficient to kill all or most of the animals of a given species, and that doses very little smaller would not kill any animals of that species. Any worker, however, accustomed to estimations of toxicity, knows that these assumptions do not represent the truth. (p. 484)

The common usage of the term "minimal competence" by educationists suggests a sense of smallest possible level of skill or knowledge at which one can still function adequately. "Minimal competence" suggests such synonomous constructions as "essential," or "highest level that is still inadequate," or "least permissible" level of skill. For example, in his explanation of mastery learning, Bloom (1968) wrote:

> The basic problem is to determine how the largest proportion of the age group can learn effectively those skills and subject matter regarded as essential for their own development in a complex society. (p. 2)

Suppose we assume a shared understanding of the meaning of the word "competence" in its noun form—without agreeing necessarily that the meaning of the adjectival form is clear—and focus on the term "minimal." "Minimal" and the noun forms "minimum" and "minimization" call to mind their opposites—"maximal," "maximum," "optimal," and the dreadful backward formation "optimize"—which are also frequently rhetorically applied to human affairs to suggest a degree of precision and determination which may not exist. ("This new plan should maximize the pay-off from our regional field staff;" "How can we minimize the flack we're likely to catch if we raise the price 10¢ a gallon? ") To speak of maximizing or minimizing some aspect of human behavior is to speak pseudo-mathematically about the natural world which does not permit the absolute treatment afforded by mathematics.

It is well to realize that many functions in mathematics and nearly all things in the natural world have no "maximum," e.g.:

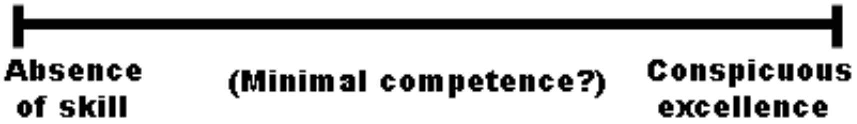
- The function $f(x) = x^{-1}$ for $x > 0$;
- The world high-jump record;
- The amount of German vocabulary (measured as number of words recognized) that a Berlitz student can acquire.

The Oxford English Dictionary gives the following unsurprising definition of "maximization": "Maximization—the action of raising to the highest possible point, position or condition." The Oxford's first illustration of the use of the word is from the works of the Utilitarian philosopher Jeremy Bentham: "The maximization of the happiness of the greatest number (1802, *Principles of Judicial Procedure*)."

It is significant that one of the earliest applications of mathematical language to human affairs should have been by the founder of Utilitarianism. For the acknowledged weakness in Utilitarianism is that it rests upon the notion of a social calculus which, in fact, does not exist. There exists no "utile" as a unit of measurement of happiness or well-being; there are no equations on can differentiate to maximize the happiness of the greatest number. To speak as though there are, is to speak metaphorically. The metaphor may have been valuable at one stage, but to mistake it for reality now places one in jeopardy of wasting his efforts in false precision and useless detail.

The notion of "minimal competence" is an educational concept. Educationists hope to use the concept to support an educational desire, viz., when may a teacher stop teaching a child because he has attained the minimal level of skill that he needs (to go to college, be a citizen, be promoted to the next grade, etc.)? In this respect, the idea of "minimal competence" raises the same definitional and practical problems as does the concept of a "cure" in psychotherapy. When is the psychotherapist's client cured so that he can leave therapy? That this question has never been given a satisfactory answer by psychotherapists ought to alert the educationist to the potential difficulty of answering the question, "When is a student minimally educated?" We suspect that 99% of all therapies are terminated not because the therapist certifies the client as "healthy," but because the client (a) graduates or changes schools, (b) runs out of money, (c) obtains a divorce, new job, face-lift, etc., or (d) grows tired of talking to the therapist or *vice versa*.

For most skills and performances, one can reasonably imagine a continuum stretching from "absence of the skill" to "conspicuous excellence":



But, it does not follow from the ability to recognize the utter absence of the skill (e.g., this paraplegic can type zero words per minute at 0% accuracy) that one can recognize the highest level of skill below which the person will not be able to succeed (in life, at the next level of schooling, or in his chosen trade). What is the minimum level of skill required in this society to be a citizen, parent, carpenter, college professor, key-punch operator? If anyone would dare to specify the highest level of reading performance below which no person could succeed in life as a parent, counter-examples of persons whose reading performance is below the "minimal" level yet who are regarded as successful parents could be supplied in abundance. And the situation is no different with a secretary or electrician --in case one wished to argue that minimal competence levels are possible for "training," if not for "educating." What is the lowest level of proficiency at which a person can type and still be employed as a secretary? Any typing rate above the trivial zero-point will admit exceptions; and if one were forced nonetheless to specify a minimal level, the rate of exceptions that was tolerable would be an arbitrary judgment.

Greenbaum (1976) alluded to an observation by Alfred Garvin to the effect that mastery or minimal competence criteria may be impossible to determine in "subject-matter areas:" "It may well be as Alfred Garvin has suggested, that performance criteria in some subject-matter areas cannot be established, since no specific extra-classroom performance is required in these areas. (p. 87)" This comment suggests a distinction roughly between what some speak of as training versus educating. But this is largley a distinction without an essential difference, and I doubt seriously that the objections I have raised about the logic of "mastery" and "minimal competence" can be answered by shifting the area of discourse from chemistry to driver training or from English literature to instrumental music.

The idea of minimal competence is bad logic and even worse psychology. Recently, in a discussion of these ideas, John Tukey wrapped up a tidy demurrer from the "minimal competence-essential skills" position in this pithy epigram: "Life is like a Double-crostic; we can do far more than we know." When one first reads the definitions of the words in the Double-crostic, he discovers that he knows only a half-dozen or so from among fifty or sixty. But eventually, through the complex and interlocking system of semantic and linguistic clues of the puzzle, all of the words and the quotation are identified. Who would be so foolish as to suggest a minimal number of definitions one must know on the first pass through the Double-crostic before the puzzle can eventually be solved?

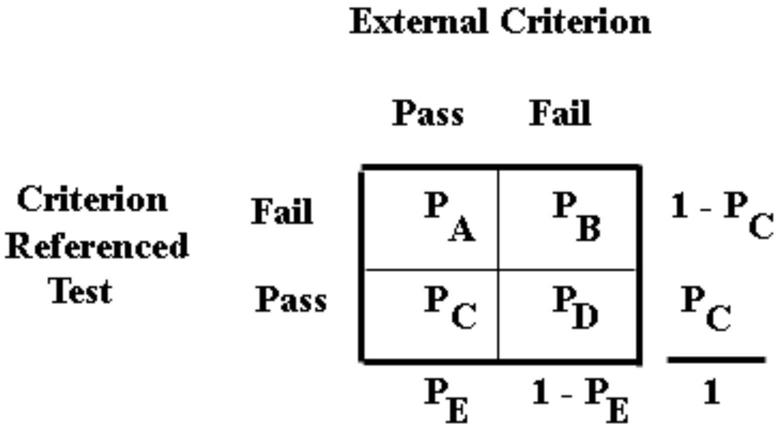
The attempt to base criterion scores on a concept of minimal competence fails for two reasons: (1) it has virtually no foundation in psychology; (2) when its arbitrariness is granted but judges attempt nonetheless to specify minimal competence, they disagree wildly.

Decision-Theoretic Approaches

The mathematical possibilities of criterion-referenced testing have not been overlooked. With characteristic fecundity, statisticians have written numerous articles on the psychometric accuracy of the criterion score on criterion-referenced tests (Kefer and Bramble, 1974; Huynh, 1976; Swainathan, Hambleton, and Algina, 1974, 1975), the reliability and validity of criterion-referenced tests, and the maximization of benefit-loss ratios through classification of examinees with criterion-referenced tests (Besel, 1973; Emrick, 1971; Hambleton and Novick, 1973; Kriewall, 1969).

Without exception, these investigators accept a "cut-off score" interpretation of criterion-referenced testing. They eschew questions of how any particular "criterion score" is justified or how it is selected. Rather, they proceed from the point at which someone (teacher, parent, school board member, or whoever) has determined a criterion score, C_x. Hambleton and Novick's (1973) treatment of the problem is illustrative: "The primary problem in the new instructional models, ..., is one of determining if $\mathbf{t}_i$, the student's true mastery level, is greater than a specified standard, $\mathbf{t}_o$." (p. 163) (In fairness to Hambleton and Novick (1973), in the paragraph preceding this quotation, they worry more than others about the arbitrariness of the criterion score in criterion-referenced testing; although in the end, they accept this arbitrariness and plunge forward.)

The problem addressed with decision theory techniques by these investigators is of the following general form: Persons are categorized into two classes on some external criterion of principal interest, e.g., "graduated vs. not graduated from college," "hired vs. not hired by an employer." The proportions of persons in these two states are denoted by P_E and $1 - P_E$. If these same persons were administered a criterion-referenced test in advance and a criterion score C_x was established by which persons would be classified as "pass" or "fail," then four combinations of passing or failing the criterion referenced test and external criterion are possible. The probabilities of persons being in each of these states can be denoted as follows:



		External Criterion		
		Pass	Fail	
Criterion Referenced Test	Fail	P_A	P_B	$1 - P_C$
	Pass	P_C	P_D	P_C
		P_E	$1 - P_E$	1

Figure 3. Display of probabilities

P_A denotes the proportion of "false negatives," i.e., persons who fail the criterion-referenced test but "pass" the external criterion. $P_D$denotes the proportion of "false positives."

The cut-off score on the external criterion is conveniently assumed to be "given" and not subject to change. In the decision-theoretic treatment of the problem, the criterion score on the criterion-referenced test is allowed to vary with the result that P_A,..., P_D and P_C will vary as a result. Clearly, it is possible to manufacture some aggregate function of good (P_B and P_C) and bad (P_A and P_D) consequences of setting the criterion score, C_x, and attempt to minimize (or maximize) one's grief (or contentment). For example, one might minimize:

$$f(C_x) = \frac{P_A + P_D}{P_B + P_C}$$

If a minimum existed, the "rational" criterion score would seem to have been found. However, this construction of the problem is highly arbitrary in that it assumes that the costs of false positives and false negatives are the same. If, on the contrary, failing persons on the criterion-referenced test who would have passed the external criterion has cost **A** and passing persons who would fail the external criterion has cost **B**, then the proper function to minimize by choice of $\mathbf{C_x}$ is:

$$f(C_x) = \frac{A(P_A) + P_D}{B(P_B) + P_C}$$

This function is clearly sensitive to the values of **A** and **B**, which would have to be determined by judgment and which would undoubtedly vary considerably depending on who assigned values to them. Assigning numbers to and would involve, for example, answering a question such as "What is the dollar cost of passing a student on this criterion-referenced test who will eventually drop out of college versus the dollar cost of failing a student on this criterion-referenced test who would eventually have graduated from college?" Thus, the arbitrariness in this technique of setting a criterion score is not encountered until the final stage, but inevitably, it is encountered.

The psychometrically trained reader will recognize that the decision-theory statement of the criterion-referenced test cut-off score problem is a special case of personnel selection theory, as explicated most fully by Cronbach and Gleser (1965).

In my opinion, all of those who have dealt statistically and psychometrically with the problems of criterion-referenced testing are guilty of misdirected precision and axiomatization. The interesting questions about criterion-referenced testing are "Whence comes C_x?" "How is one criterion score justified over another?" The decision theory and psychometric questions are routine, and standard techniques have merely been clothed in the language of criterion-referenced testing and offered as answers. The answers are correct and valid given the premises. But the entire endeavor (viz., to treat criterion-referenced testing statistically and psychometrically) has been undertaken without a sense of the critical concern. Of what concern is it that *n* items must be sampled or a cut-off score set at C_x to minimize false negatives, if at the very bottom of it all the decision to "pass' 30% vs. 80% is judgmental, capricious, and essentially unexamined?

"Operations Research" Methods

This technique for setting a criterion score is so named because it is based on the general approach of operations research of maximizing a valued commodity by finding an optimum point on a mathematical curve or a graph. An illustration will clarify this meaning.

Taking his cue from Bormuth's (1971) application of operations research strategy to determining optimal "readability" of instructional passages, Block (1972) presented a method which was alleged to be the rationally justifiable technique for determining the criterion score on criterion-referenced tests. Theoretically, the researcher would teach many different randomly equivalent groups until they achieved various levels of proficiency on a "criterion-referenced test," e.g., 10%, 15%, 20%, ..., 95%, 100%. Furthermore, all of the groups would be measured on an external measure of valued outcomes, e.g., performance on a retention or transfer of learning scale, income at age 40, "life success," etc. Next, a graph relating the degree of mastery on the criterion-referenced test and performance on the valued outcome scale is drawn (see Figure 4).

That level of performance on the criterion-referenced test for which the valued outcome score is maximized becomes the "rationally" determined criterion score. It is immediately clear that this method does not satisfactorily resolve the criterion-score determination problem unless the curve in Figure 4 is non-monotonic, i.e., unless at some point between 0% and 100% it bends and starts to return to the baseline of the graph. For unless this bend occurs, the criterion score on the criterion-referenced test which maximizes the valued outcome will be 100%—an impossible level of perfection and a trivial and thoughtless standard.

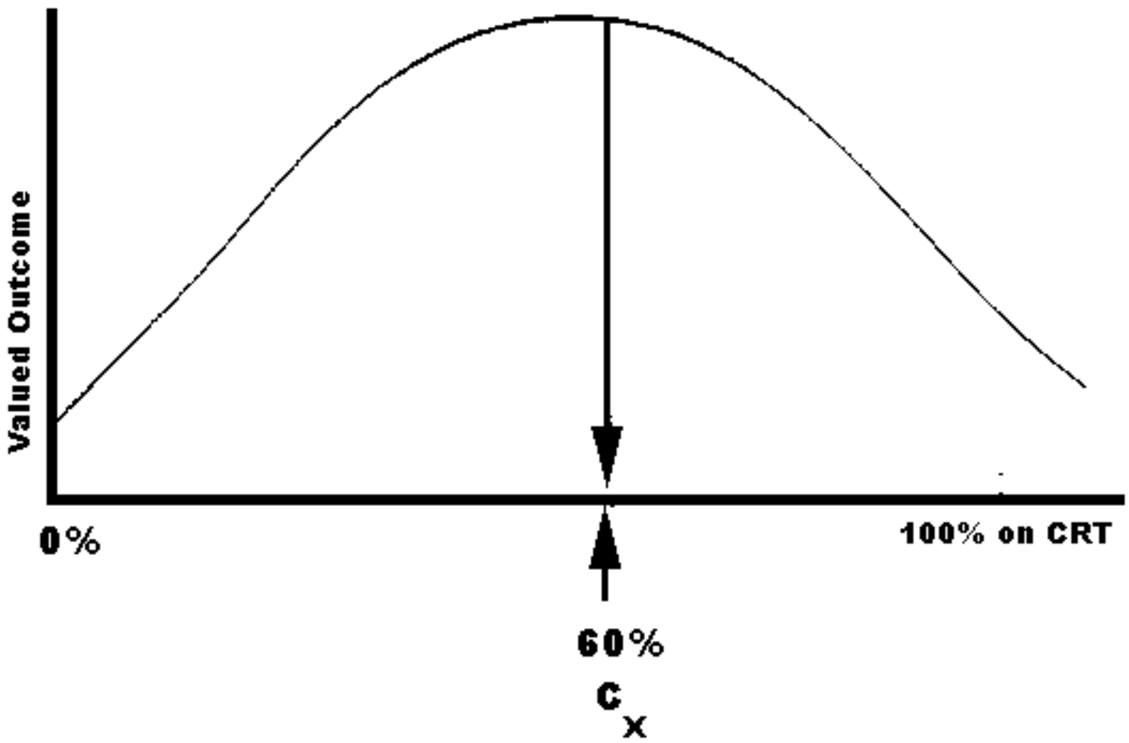


Figure 4. Hypothetical relationship between a criterion-referenced test and a valued outcome.

I suspect that non-monotonic graphs will be rare exceptions when both the criterion-referenced test and the valued outcome scale are measures of cognitive performance. That is, I expect performance on the valued outcome scale to increase monotonically as performance on the criterion-referenced test increases.

One way around this problem is to introduce a second valued outcome that bears an inverse relationship to degree of mastery on the criterion-referenced test, e.g., interest or attitude toward the topic learned, and students develop poorer attitudes the longer they study the topic. Consider the graphs in Figure 5.

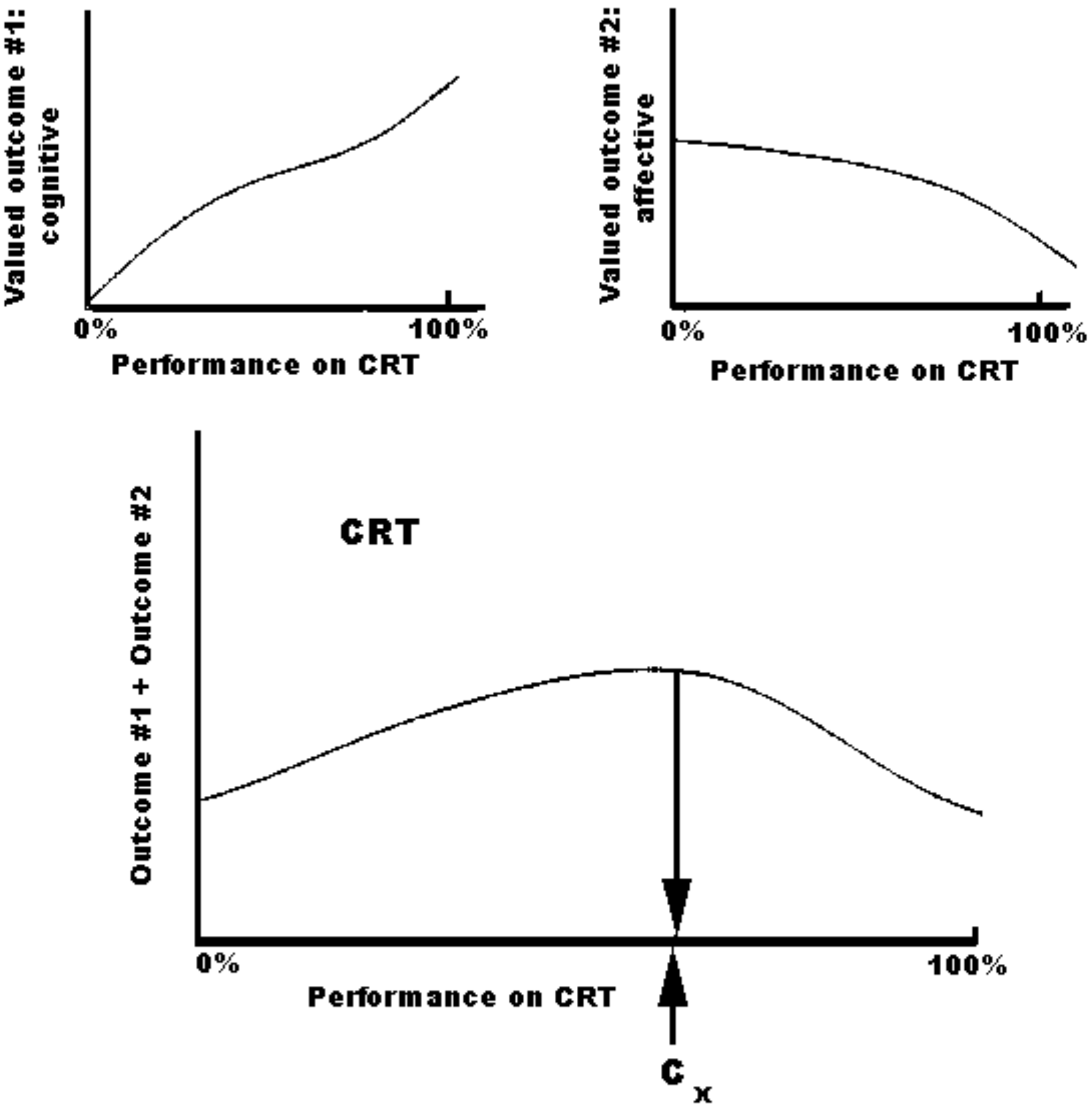


Figure 5. Separate and composite relationships between a criterion-referenced test and two valued outcomes.

Under the conditions in Figure 5, a single "criterion score" can be found for which the composite outcome (1 + 2) is maximized. This would seem to provide a "rational" and non-arbitrary method of setting a criterion score on the criterion-referenced test. But, the elimination of arbitrariness is illusory. The unreliable judgmental element in this method is hidden in the composite outcome. To weight the cognitive and affective outcomes equally in forming the composite is an arbitrary choice of composites from among the following general class of composites:

Composite Outcome = a(Outcome$_1$) + b(Outcome$_2$)

This arbitrariness is closely comparable to the problem of weighting false positives and false negatives in the "decision-theory" approach.

There is little on the face of the problem to recommend the composite "outcome$_1$ + outcome$_2$" over the composite "2 x outcome$_1$ + outcome$_2$." This latter composite would substantially shift the "criterion score" in Figure 5 to the right along the baseline. The only empirical attempt to set criterion scores by the "operations research" method resulted in precisely this ambiguity.

The results of Block's (1972) empirical study appear in Figure 6. Ninety-one eighth-grade students were taught matrix algebra. The subjects were nearly equally assigned to five groups: Control, 65%, 75%, 85%, and 95% mastery as measured by a criterion-referenced test. In the four "percentage mastery" groups, students were taught and reviewed the lesson until they could answer correctly the designated percentage of questions on the mastery test; the control group simply studied the lesson and took the mastery test. A "valued outcome" criterion measure was developed; it included twenty items. This external test was administered after all the subjects had been taught up to or exceeding their group's designated mastery level. Secondly, an "attitude toward algebra" scale was administered at the completion of the study. The measures on the external achievement test and the attitude scale for the five mastery level groups appear as Figure 6.

One first inspects Figure 6 for any evidence of non-monotonic relationships. Although Block made much of the bend in the "attitude" curve in Figure 6, the relationship between the criterion-referenced test and the attitude scale is not convincingly curvilinear. (Since group variances were not reported by Block, I was unable to carry out an exact test of departure from linearity on these data. However, by solving his F-ratio backwards, I determined that the standard deviation of scores on the attitude scale was about 5. Using this estimate of standard deviation, the F-ratio for the quadratic component in a trend analysis of the means is only 2.52 which fails to reach even the 90th percentile (2.84) in the F-distribution with 1 and 68 degrees of freedom.

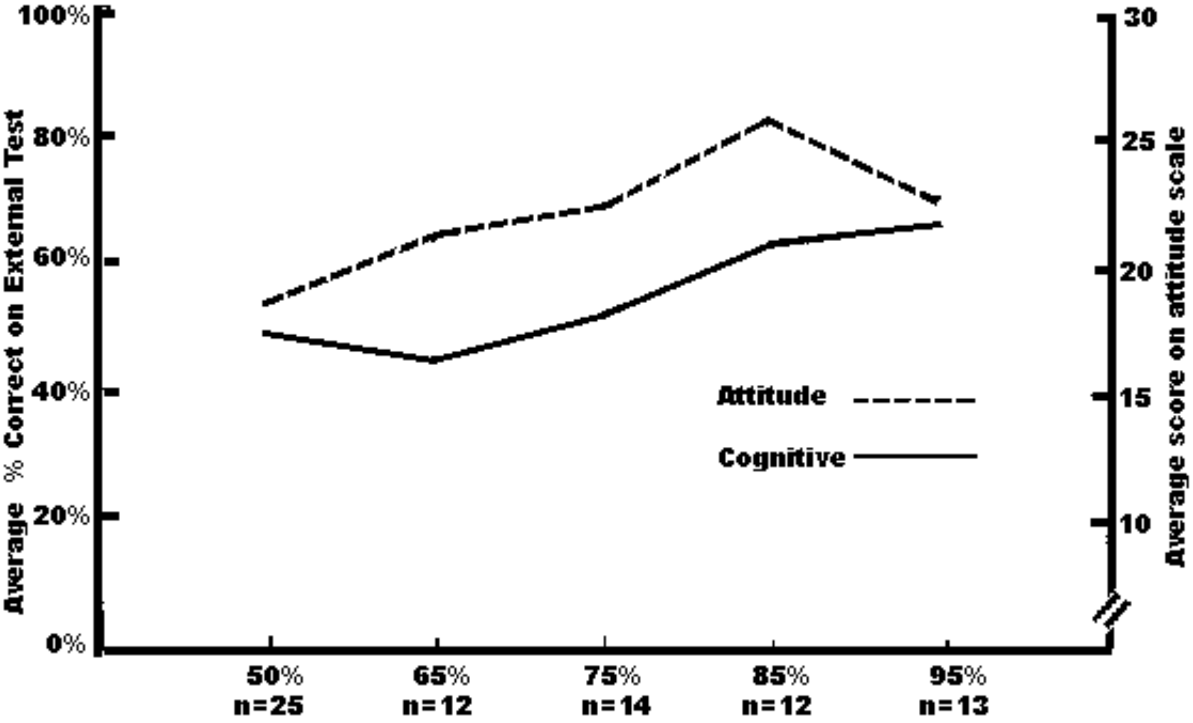


Figure 6. Attitude and Achievement for five mastery level groups. (Note: After Block, 1972.)

The "achievement" curve is definitely monotonic, as was expected. Block (1972) concluded:

> ... maintenance of the 95 percent level (of mastery) best maximized the learning represented by the cognitive criteria while maintenance of the 85 percent level best maximized the learning represented by the affective criteria. Given a model for relating scores on the cognitive criteria to scores on the affective criteria, therefore, it would have been possible to set a mastery standard for the algebra sequence. (p. 14)

Even if one accepts the tenuous evidence for non-monotonicity for the "attitude" curve in Figure 6, one is left with the problem of the arbitrary weighting of achievement and attitude in a composite outcome before the "criterion score" can be determined. This application of the "operations research" approach has reduced the appearance, but not the essence, of arbitrary judgment in setting a criterion score.

There exists a weaker form of the "operations research" argument for determining a criterion score. Suppose that beyond some point of proficiency on a criterion-referenced test, one achieves no gains on an external valued outcome (as in Figure 7).

Then, a persuasive argument could be advanced to the effect that the point C_x on the criterion-referenced test scale represented a "mastery point," in the sense that once a pupil reached C_x, one ought to stop teaching him since greater proficiency on the criterion-referenced test brings zero returns on the external valued outcome. Such reasoning has practical value to the extent that one encounters the general form of curve depicted in Figure 7, viz., a curve with an abrupt bend or corner. One is unlikely to encounter such psychometric anomalies.

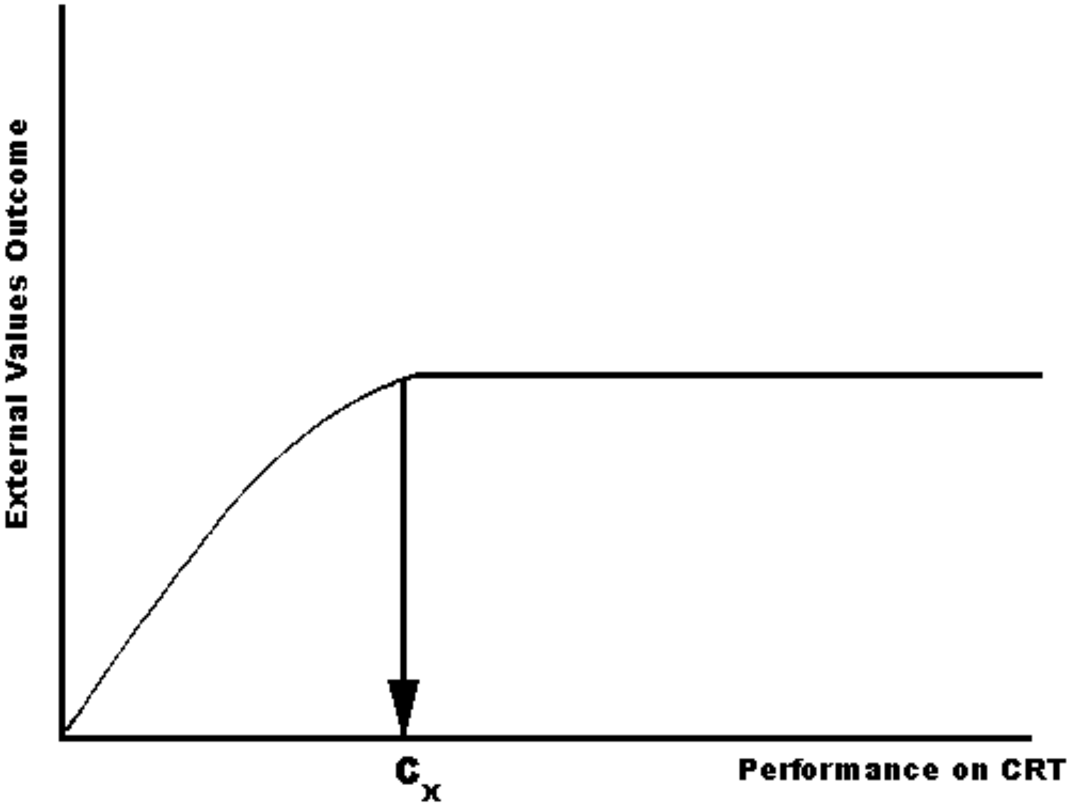


Figure 7. Hypothetical relationship between criterion-referenced test and a valued outcome.

Robert Glaser and John Tukey (personal communications, 1976) pointed out some psychological phenomena which suggest that some modification of this line of thinking might be useful. Glaser noted that young children, ages 6-8, can be trained to about 70% accuracy in single-digit addition. Training beyond that level fails to yield any greater accuracy; improved accuracy comes only with age. Tukey reported that no matter how intensively one trains telephone operators, they never become more than 98% accurate. Both examples suggest some psycho-physical limit of attention and human accuracy. Could these values serve as natural criterion levels? Is 70% a psychologically justified criterion score for an eight-year old child on a test of single digit addition? Is 98% a rationally justified criterion level for the training of telephone operators? As enticing as these instances appear, on closer examination their implications for instructional decision-making and testing will become, at once, complex and unclear.

Living Without "Standards"

In a recent conversation on the question of setting criterion scores on tests, Michael

Scriven acknowledged the distressing arbitrariness of setting standards by existing methods. But, he went on to voice the position that something was better than nothing, i.e., that the injustice and ineptitude resulting from an absence of standards is worse than the consequences of adopting arbitrary ones. Emrick (1971) espoused the same position in writing:

> It is not difficult to show that the traditional measurement procedures are inadequate, or at best arbitrary as a method of identifying student skill mastery. For example, using criterion-referenced procedures IPI has suggested an 85% correct minimum as a mastery criteria for any skill test (of which there are over 400). Although this criteria does have intuitive appeal, there is no convenient analytical or empirical justification for it. Just as various skills may differ in level of difficulty in terms of mastery, so also might the optimal performance criteria in the test situation vary. It may easily be that for some skills, a test score of 60% is indicative of mastery, whereas, for others a score of 90% or higher would be required. In short, the issue is not whether a criterion-referenced testing Procedure is or is not appropriate to IPI, but rather how and at what level each criterion should be set. (p. 321)

In the last sentence of the quotation, Emrick states the nub of the argument with which I take exception. The more general question suggested by Emrick's claim is one I cannot address adequately here. However, it ought to be commended to the attention of educational philosophers and empiricists. "Is any increase in precision—in stating behaviorally what one wishes to teach, in quantifying decisions now made less formally —an unconditional good, which, though it may not necessarily represent a gain in value, surely cannot be worse than imprecision. ("An educated man demands no more exactness than is allowed by the subject-matter being dealt with." [Kaplan, 1964, p. 283.]) To ask for greater precision than the circumstances permit is foolish, and it may be imprudent as well. The issue, as I see it, is precisely whether a criterion-referenced testing procedure entailing criterion or mastery levels is appropriate. I think not. With respect to setting criterion scores on criterion-referenced tests, nothing may be safer and better than an arbitrary something.

To my knowledge, every attempt to derive a criterion score is either blatantly arbitrary or derives from a set of arbitrary premises. But arbitrariness is no bogeyman, and one ought not to shrink from a necessary task because it involves arbitrary decisions. However, arbitrary decisions often entail substantial risks of disruption and dislocation. Less arbitrariness is safer.

Consider a pertinent actual example. A large school district in Florida in the summer of 1975 decided to reexamine its definition of "mentally retarded." One imagines that their motives originated both in the Zeitgeist for "mainstreaming" and in the public concern about overuse of the "mentally retarded" label. The administrators in the district decided to substitute a new definition of "mentally retarded" (which had been established by the American Association for Mental Deficiency) for the old definition of "IQ below 75." The new AAMD standard for "mentally retarded" involved a conjunction of several indicators each with an arbitrary cut-off point. (It is probably safe to say that it was put together around a committee table with little idea of what percentage of the school population would thereby be designated "mentally retarded.") A1though it is to be expected that the percentage of persons simultaneously below cut-off scores on several even moderately correlated variables is extremely small, the school district personnel were unpleasantly surprised in September 1975 when there was a mass emptying of pupils from the mentally retarded into the regular classes. Regular classrooms were inundated with erstwhile "mentally retarded" pupils for whom teachers had neither training, experience, nor materials. The first month of school was chaotic. Then the administration rescinded the order, and the old definition of mental retardation was reinstated.

The whole matter might have been dealt with more intelligently and less arbitrarily. The concern with which the administration attempted to deal was that too many pupils—about 10%—were being classified as "mentally retarded" by the "IQ below 75" definition. The less disruptive course would have been to plan to change the percentage of pupils in mentally retarded classes from 10% to 8% or 7%—by either lowering the cut-off slightly on the IQ test or asking special education teachers to nominate the best prospects for integrating into regular classrooms—see how the system responded to this change, and proceed.

In this example lies a means of coping with the standards problem. Perhaps the only criterion that is safe and convincing in education is change. Increases in cognitive performance are generally regarded as good, decreases as bad. Although one cannot make satisfactory absolute judgments of performance (Is this level of reading performance good or masterful?), one can readily judge an improvement in performance as good and a decline as bad.

My position on this matter is justified by appeal to a more general methodological question in evaluation. Is all meaningful evaluation comparative? Or do there exist absolute standards of value? I feel that in education there are virtually no absolute standards of value. "Goodness" and "badness" must be replaced by the essentially comparative concepts of "better" and "worse." (In the same conversation alluded to above, Michael Scriven was asked whether he believed that all evaluation was necessarily comparative. He answered, "No, only all good evaluation is comparative.") Absolute evaluation in education—as reflected in such endeavors as school accreditation and professional licensing—has been capricious and authoritarian. On the other hand, the value judgments based on comparative evidence impress us as cogent and fair. Data from comparative experiments, norm-referenced tests and longitudinal assessments of change are comparative evidence, and thus enjoy a presumptive superiority over non-comparative evidence. The economist Kenneth Boulding (1953) made the same point about social systems in general: "Almost everybody is sensitive to comparative statistics. It is often not the absolute value of a variable which is significant but the difference between your value and that of some other comparable person or organization" (p. xxxii).

I am confident that the only sensible interpretations of data from assessment programs will be based solely on whether the rate of performance goes up or down. Interpretations and decisions based on absolute levels on performance on exercises will be largely meaningless, since these absolute levels vary unaccountably with exercise content and difficulty, since judges will disagree wildly on the question of what consequences ought to ensue from the same absolute level of performance, and since there is no way to relate absolute levels of performance on exercises to success on the job, at higher levels of schooling, or in life. Setting performance standards on tests and exercises by known methods is a waste of time or worse.

In education, one can recognize improvement and decay, but one cannot make cogent absolute judgments of good and bad. It is well to recognize that in proposing "change" as the solution to the standards problem, one introduces a problem of standards—or absolute judgment—at a second level, viz., How much increase is good or sufficient? How much loss is tolerated before action is taken? Here one confronts precisely the problem of a criterion score—how many percentage points decline can be tolerated?—which was avoided by substituting the criterion of change for an absolute criterion score. But the substitution was not futile. Considerable clarity and consensus were bought when "change" was substituted for "absolute level of performance," even if all problems were not solved.

Acknowledgment

This monograph was prepared under a grant from the Carnegie Corporation. Mary Lee Smith contributed at every stage of the preparation of this monograph; reading the literature, developing the ideas, outlining and editing the final product. Nancy W. Burton helped in developing the topic. Laura A. Driscoll and Gregory C. Camilli deserve thanks for assistance in literature search and data analysis. Conversation among Glass, Smith, Burton, and Robert Glaser on a day in June 1976 got things started; we are much indebted to Glaser for helping set things off in a direction that now pleases us.

References

Andrews, B.J. & Hecht, J.T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36*, 45-50.

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement*. (2nd ed.) Washington, DC: American Council on Education.

Besel, R. (1973). Using group performance to interpret individual responses to criterion-referenced tests. Paper presented at Annual Meeting of the American Educational Research Association, New Orleans.

Block, J.H. (1972). Student evaluation: Toward the setting of mastery performance standards. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Bloom, B.S. (1968). Learning for mastery, *Evaluation Comment, Vol. I*, No. 2.

Bormuth, J.R. (1971). Development of standards of readability: Toward a rational criterion of passage performance. Final report, U.S. Office of Education, Project No. 9-0237. Chicago: University of Chicago.

Boulding, K.E. (1953). *The organizational revolution*. New York: Harper and Brothers.

Cronbach, L.J. & Gleser, G.C. (1965). *Psychological tests and personnel decisions*. (2nd ed.) Urbana, IL: University of Illinois Press.

Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Emrick, J.A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement, 8*, 321-326.

Fair, J. (1975). *National assessment and social studies education: a review of assessments in citizenship and social studies by the national council for the social studies*. U.S. Government Printing Office.

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

Glaser, R. & Klaus, D.J. (1962). Proficiency measurement: assessing human performance. Pp. 419-474 in Gagne, R.M. (Ed.), *Psychological Principles in Systems Development*. New York: Holt, Rinehart, and Winston, 1962.

Glaser, R. & Nitko, A.J. (1971). Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational Measurement*. Washington, DC: American Council on Education, 1971, 625-670.

Greenbaum, A. (1976). *A study of the national assessment*. A book produced under a grant from the Carnegie Corporation. In press, 1976.

Hambleton, R. K. & Novick, M.R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*, 159-170.

Harris, M.L. & Stewart, D.M. (1971). Application of classical strategies to criterion-referenced test construction. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika, 41*, 65-78.

Ivens, S.H. (1970). An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Doctoral dissertation, Florida State University.

Jackson, R. (1970). Developing criterion-referenced tests. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1970. (ERIC Document No. ED 041 052).

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco: Chandler.

Kifer, E. & Bramble, W. (1974). The calibration of a criterion-referenced test. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document No. ED 091 434)

Kriewall, R.E.(1969). Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Doctoral dissertation, University of Wisconsin. Ann Arbor, MI: University Microfilms, No. 69-22417.

Lindvall, C.M. & Nitko, A.J. (1975). *Measuring pupil achievement and aptitude*. (2nd ed.) New York: Harcourt, Brace, and Jovanovich, 1975.

Mager, R.F. (1962). *Preparing instructional objectives*. Palo Alto, CA: Feardon Publishers, 1962.

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

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

Popham, W.J. (1973). *Establishing performance standards*. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J. & Husek, T.R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1-9.

Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of criterion-referenced tests: a decision-theoretic formulation. *Journal of Educational Measurement, 11*, 263-267.

Swaminathan, H., Hambleton, R.K., & Alaina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement, 12*, 87-98.

Trevan, J.W. (1927). The error of determination of toxicity. *Proceedings of the Royal Society, Series B, Vol. 101*, 483-514.

Tyler, R.W. (1973). Testing for accountability. In A.C. Ornstein (Ed.) *Accountability for teachers and school-administrators*. Belmont, CA: Feardon Publishers.

Wiersma, W. & Jurs, S.G. (1976). *Evaluation of instruction in individually guided education*. Reading, MA: Addison-Wesley, 1976.