Standard Setting Issues and Methods

Lorrie Shepard
University of Colorado

Previous methodological reviews and the controversy regarding the adequacy of standard-setting technology are summarized. The judgmental nature of all standard-setting methods is examined, and the debate about whether fallible standards are better than none is recast in the context of three different test uses: pupil diagnosis, pupil certification (for high school graduation or professional licensure), and program evaluation. Exemplary standard-setting methods are reviewed, representing the following major approaches: (1) judgments of test content; (2) judgments about mastery-nonmastery groups; (3) norms and passing rates; (4) empirical methods for discovering standards; and (5) empirical methods for adjusting cutoff scores, given a standard on an external criterion measure. Standards based on the performance of judged mastery groups (the Contrasting Groups method) and certain uses of normative data are likened to Known Groups validation. Recommendations are made for selecting standard-setting techniques depending on test use, including pupil diagnosis, pupil certification, and program evaluation. Future research on standard setting is discussed in the context of improving practical aspects of judgmental methods.

For some time after criterion-referenced testing was invented (Glaser, 1963), standards or passing scores on the test were considered to be the distinguishing characteristic of criterion-referenced tests. These were the criterion levels required to make absolute rather than relative interpretations of performance. Now that the original intent of criterion referencing is more clearly understood to mean detailed specification of the content or behavior domains to be assessed (Millman, 1974; Popham, 1975), it is acknowledged that standards are not always needed for applications of criterion-referenced tests.

Hambleton, Swaminathan, Algina, and Coulson (1978) distinguished two uses for criterion-referenced tests: (1) to estimate examinee domain scores and (2) to assign examinees to mastery states. The first use requires a report of how much of a content domain an examinee knows; the second use involves comparing the examinee's score to a standard to determine if he or she knows enough to be considered a master. Obviously, only the second use requires a judgment about how much knowledge constitutes mastery.

When mastery classifications are desired, there is no ready technical solution to the setting of standards because there are no universal standards, and whether reasonable standards can be set at

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 4 Fall 1980 pp. 447–467

all is a matter of some controversy. The purpose of this paper is to summarize the issues involved in judging the adequacy of standard-setting methods for different applications of criterion-referenced testing. The evaluation of differences in methods provides some insight for selecting the preferred approach in a particular situation. For some purposes, the use of more than one method is recommended to compensate for the limitations in each method. More importantly, however, the case is made that for some uses of criterion-referenced tests, better interpretations of test results can be made entirely without the imposition of inaccurate standards.

Previous Reviews

Several reviews of the standard-setting literature have already been written (Glass, 1978b; Hambleton & Eignor, 1979; Hambleton, Powell, & Eignor, 1979; Jaeger, 1979; Meskauskas, 1976; Millman, 1973; Shepard, 1980), making a new literature review almost superfluous. With this in mind, the descriptions of methods presented in this article will be as brief as possible. Instead, several very different categories of test use will be described and the way the characteristics of these uses influence the validity and practicality of different standard-setting methods will be considered.

Controversy over Standards

Glass (1978b) started the controversy over the adequacy of standard-setting methodology. He examined various techniques for arriving at cutoff scores and concluded that all are arbitrary or are built on arbitrary premises. The logical flaws he identified in specific methods and the contradictions between methods led him to the conclusion that standard setting should be avoided and that other ways should be found to attach value to testing results.[1]

The rebuttal that has been most widely adopted in response to Glass, that standard setting may be judgmental but it need not be capricious, was given by Popham (1978). All writers in the field acknowledge that standard-setting techniques are judgmental (Block, 1978; Hambleton, 1978; Hambleton & Eignor, 1979; Jaeger, 1976; Shepard, 1976, 1979); but Popham (1978) argued that this does not prevent educators from arriving at reasonable and defensible standards for what they believe are acceptable levels of performance. The debate, therefore, has centered around two meanings for the word arbitrary, only one of which connotes thoughtless and whimsical decisions.

As different standard-setting approaches are examined in this paper, it is apparent that they are all techniques for gathering information and for drawing attention to the choices to be made, so that the necessarily arbitrary decisions will be as considered and as logical as possible. However, just because standard setting can be done carefully does not mean that Glass's (1978b) criticisms have been refuted. The most important point, which will influence the choice about whether or not to set standards, is that there is always error attached to the selection of cutoff scores. Individuals immediately on either side of the standard will be virtually indistinguishable from one another. With a good test, valid distinctions can be made between those who are well above or well below the standard; but pass-fail distinctions near the cutoff will have poor validity because a continuum of performance has been "arbitrarily" dichotomized.

---

[1]As an historical note, it is worth mentioning that Glass's work was initially prompted by a grant to address the use of standards for interpreting large-scale assessment results and was expanded to consider the issue of minimum competency testing; although Glass did not share these delimitations with his audience, it is clear that his focus was not classroom level uses of criterion-referenced tests.

In addition to the assurance that standard setting can be reasonable rather than capricious, several authors have also replied to Glass that "potentially flawed standards are better than none" (Hambleton, 1978; Popham, 1978; Scriven, 1978). If pass-fail decisions are inevitable, good test information, even with an arbitrary cutoff score, will lead to better decisions than those that would be made without the test. This is especially true if the decisions made in the absence of the test are equally arbitrary but less well considered. There is also the belief that the very presence of the test hurdle will so motivate learning that quality will be ensured even when very few appear to fail the test. These assertions are compelling but may or may not be true in particular applications. Such arguments depend first upon the validity of the test to inform a specific decision. Validity issues in criterion-referenced testing are outside the province of this paper but are covered in Linn (1979) and Shepard (1980). In addition, to conclude that imperfect standards are better than none, there must be some evidence that, indeed, pass-fail classifications will be made with or without the test and that the salutary effects of having the standards in place are substantial enough to offset the costs of classification errors. It is part of the intent of this paper to restage the debate about standards in the context of different test uses to determine whether the benefits of standards outweigh the costs of arbitrary classifications.

USES OF CRITERION-REFERENCED TESTS

It is generally understood that validity is not an inherent and fixed attribute of a test; rather, validity will depend on how a test is used (American Psychological Association, 1974; Cronbach, 1971). It is not a new idea, then, to say that technical issues including standard setting will be influenced by the purpose of testing.

Perhaps the most global categorization of test uses is the distinction between individual and group interpretations of test results. This important difference corresponds to the two purposes identified by Cronbach (1970), selection and classification of persons versus evaluation of treatments. The differentiation between individual and group level interpretations is adopted here because it so often has implications for the level of technical accuracy required. In addition, for criterion-referenced tests this distinction makes a difference in whether cutoff scores will be essential for making the intended educational interpretations or decisions, i.e., standards may be unavoidable to make pass-fail decisions for individuals, but there are better ways to evaluate the performance of schools or educational programs. The level of the decision will make a difference with regard to standard setting.

The individual level of test use can be divided further into two major categories. Pupil classification for instructional purposes, which might also be called "individual diagnosis," is distinguished from an "individual certification" test use. The latter is intended to prove the examinee's level of accomplishment to an external and probably more skeptical, less nurturant audience. These two kinds of individual level uses are analagous to formative and summative evaluation (Scriven, 1967). The former involves judgment of performance but is focused on improvement; the latter renders a final judgment of quality. The two types of individual test use also differ in the proximity of the test to the instructional process. Tests used for individual diagnosis are a part of everyday instruction; certification tests are more removed and are intended to verify learning outcomes. These two types of individual use plus program evaluation purposes are described further below.

Pupil Diagnosis

Pupil diagnosis refers to classroom level decisions about an individual pupil. It is the use that Glaser (1963) had in mind when he first advocated criterion-referenced testing. In order to individual-

ize instruction, and hence to help each student learn as much as he or she is able to, there must be a clearly defined progression of educational objectives and tests that accurately report a student's level of performance. Cutoff scores on the tests indicate when a student has learned one topic and is ready to go on to the next. This type of use requires extensive pools of test items matched to the objectives that define a specific curriculum and that can be administered at the discretion of the classroom teacher to make short-term instructional decisions about needed review, workbook assignments, reading group placement, advancement to new material or remedial tutoring. It is implausible that such tests could be administered at fixed times on a large scale, for example, state-wide. To provide useful diagnostic information, the tests have to be tailored to a specific curriculum and administered just at the time when the teacher is uncertain about what to do next with a particular child.

Pupil Certification

Unlike diagnostic tests that cover specific instructional objectives, certification tests are removed from the teaching-learning process and must be comprehensive. Professional licensure examinations and minimum competency tests for high school graduation are examples of certification uses of criterion-referenced tests. Annual tests for grade-to-grade promotion may be in between both diagnostic and certification purposes, since pass-fail decisions would be linked to remediation. However, this use still seems closer to "certification," since all students must pass the same hurdle at the same time and any instructional response for remediation would be delayed rather than immediate.

Certification tests are usually administered by an external agency, not by the classroom teacher. They are given to confirm a pupil's knowledge, even though the pupil might already have obtained course grades or classroom placements that imply mastery. Certification tests are not given primarily for the benefit of the test taker. Instead, credentialling exams protect some larger public by only certifying those who score well enough to be considered competent.

Program Evaluation

Program evaluation is a broad term for a variety of group level decisions. Judgments about quality may be directed at the educational program of a particular school, a district-wide curriculum, or a specially funded project. In each case the achievement of a group of students is studied to determine the effectiveness of instruction. At least with regard to its effect on standard-setting methods, the category of program evaluation can be interpreted loosely enough to include state and local accountability programs and research studies. Program evaluation not only entails overall judgments but also information about relative strengths and weaknesses within a program, such as teaching math computation and math comprehension or the development of healthy attitudes as well as cognitive skills.

Unlike individual test uses, which may require dichotomous classifications of pupils, judgments about programs do not necessarily depend on making explicit pass-fail decisions about pupils. What is necessary is that the test data be aggregated and summarized appropriately and that some benchmark be used to attach value to the results. Absolute standards are one way to try to interpret the "goodness" or "badness" of aggregate performance levels. But, as will be discussed in a later section, pass-fail cutoff scores may not provide the best yardstick for judging what happens to groups at different locations on the performance continuum or for diagnosing program strengths and weaknesses.

STANDARD-SETTING METHODS

Methods which Assume Mastery Is an All-Or-None State

Meskauskas (1976) organized his review of standard-setting methods by distinguishing continuum and state models. For continuum models the ability being assessed is assumed to be continuous with a mastery region at the upper end; a discrete cutoff is sought because a dichotomous decision is needed. For state models, mastery is presumed to be all or none; that is, an examinee either has the skill or he does not.

The difficulty with state models is the implicit expectation that mastery will be recognizable as 100% performance. Meskauskas used a quotation from Davis and Diamond (1974) to illustrate the presumption of a 100% standard:

> Strictly speaking, mastery is defined as complete knowledge, skill, or control; so "partial mastery" is as self-contradictory a phrase as "partial uniqueness." The term "mastery," therefore, should be used to describe the status of only those who, it may be inferred, can mark correctly all the items in the population of which the subset that makes up a criterion-referenced test is a representative sample. (p. 133)

It is unrealistic to expect perfection, however, because of errors in the test and occasional uncharacteristic mistakes made by the examinee, like those made by competent adults in balancing a checkbook. Standard-setting methods based on a state model (Emrick, 1971; Roudabush, 1974), therefore, are simply algorithms for taking account of measurement error. Glass (1978b) termed this approach "counting backwards from 100."

State-model methods for setting standards are given short shrift in this review, then, for two reasons:

1. Except for very tiny achievement domains, such as single-digit addition problems (with vertical, numeric format), a continuously distributed trait is more plausible than all-or-none states.
2. Once it is established that some allowance should be made for measurement error, the procedures for deciding on how much to adjust the 100% standard and for weighing the two kinds of classification error are the same as those proposed for continuum models.

All of the following sections, therefore, refer to different approaches for standard setting aimed at dividing a continuum into mastery and nonmastery categories.

Methods for Dichotomizing a Continuum

In continuum models, the cutoff score on the test is chosen to reflect the least amount that an examinee can know and still be considered a master. All of the methods proposed to formalize the selection of this cutoff point are decision strageties to help in thinking about what amount of knowledge should be required. They are exercises, very much like those an instructor might use informally in trying to decide whether the cutoff score for a passing grade should be set at 60% or 65% of the cumulative points. New meaning about the percentage scores is gained by examining the test papers of the marginal students (in the range 60% to 65%) to see if they consistently know the answers to the most fundamental questions. This insight is different from those provided by a priori expectations or comparisons to students in previous years. Different standard-setting methods also provide different in-

sights. The following sections represent the major approaches for considering where the standard should be. Because the various methods have been reviewed so extensively previously, only one exemplary method is fully described in each section. Other variations of the approach are noted.

Absolute Judgments of Test Content

Criterion-referenced testing has an important connotation of absolute, rather than relative, interpretations of achievement. For this reason, the most obvious method for setting standards has been to inspect test content and to decide what percentage of correct answers looks like evidence of mastery. In this way, only the merit of the questions and the expectations of the examiners determines the standard rather than the performance of examinees.

Perhaps the most straightforward technique for inspecting test content is the Angoff (1971) approach. For this method, as well as other judgmental methods, standard setters are asked to imagine a minimally qualified individual. A mental picture of what levels of skills the just-barely-passing-candidate must have is necessary whether the test is meant to certify high school graduates or medical doctors. Examples will sometimes help in conjuring an image of a master who makes a tolerable number of errors and in distinguishing this individual from one whose performance makes nonmastery more plausible. Judges might, for example, think of mistakes they themselves make as still being consistent with mastery. But there is a point where the errors become too numerous to excuse and the judge says to himself/herself, "the individual with this score is more similar to the obvious incompetents, the illiterate, or the physician who unwittingly administers lethal drugs."

The validity of judgmentally set standards depends on the definition of the minimally qualified examinee. Discussion and training with the judges can increase the amount of thought that is given to the problem and improve the agreement among judges. Of course, there is still no way to remove either the subjectivity of this crucial definitional stage or the variability in how the definition is operationalized. The subjectivity of this process is both its strength and its weakness. By completely ignoring normative data, the judges are able to assert what should be rather than what is. At the same time, the many different standards that are produced by the different expectations of the judges make all of their separate standards defensible; hence, there are always challenges or alternatives to the final cutoff point.

Once judges have defined mastery, then, using the Angoff (1971) method, they read all the test items and assign a probability value to each one. The probability is a subjective estimate of how likely it is that the just-barely-qualified person will answer correctly. The sum of these probabilities for all the items in the test becomes the cutoff score. So, for example, if a judge thinks that a marginal master will have an 80% chance on each of 10 items, the passing score on the test would be 8.

Other methods for judging test content provide standard setters with slightly more complicated formulas for considering item difficulty. The Nedelsky (1954) method is the oldest procedure and is widely used, especially in the health professions. Judges arrive at probabilities for items indirectly by eliminating the wrong choices that they believe a minimally competent individual would clearly know are wrong and then by computing the probability of guessing correctly from the remaining choices. The Nedelsky procedure creates some practical problems because the task of casting out wrong choices is unfamiliar to most judges. Furthermore, for new competency-testing programs, where test content has already been curtailed to include only minimal skills, it may be that judges will feel that *all* the distractors should seem clearly wrong to the competent examinee. Because the Nedelsky procedure limits the probabilities to discrete steps (see Brennan & Lockwood, 1980), the resulting standard will either be the unrealistic 100% (eliminate all the wrong answers with certainty) or a very generous

50% (always guessing between the correct answer and the next best answer). Empirical studies have consistently found that the Nedelsky procedure produces lower standards than other methods based on judgments of test content (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Kleinke, 1980; Koffler, 1980; Skakun & Kling, 1980).

The Ebel (1972) method for deciding on standards is similar to the Angoff (1971) approach with the additional complexity that judges are first asked to categorize items by relevance and difficulty. Meskauskas (1976) noted that this particular categorization scheme might be awkward to use in practice because the dimensions are correlated, that is, *essential* items would have to be rated as *easy* for a group of masters. However, some sorting procedure is recommended, since consistent probabilities can then be assigned to clusters of items similar in importance and difficulty. Educational Testing Service (1976) did this, for example, by asking judges to locate items on a probability continuum when implementing the Angoff method with the National Teachers Examination.

All of the judgment methods are common sense approaches for wrestling with the standard-setting task. The Angoff (1971) procedure is favored here primarily because it is simpler. Simplicity has the advantage of not obscuring the basic subjectivity of the decisions; judges more clearly have the sense that they are "pulling the probabilities from thin air." This uneasiness is essential for dealing properly with the limited validity of the cutoff score. When more complicated formulae are used, there may be a false sense of scientific precision.

Glass's (1978b) indictment of standard-setting techniques was based not only on their subjectivity but also on the serious discrepancies in the standards they produced. Glass cited an early study by Andrew and Hecht (1976); more recent studies, comparing different combinations of the Angoff, Ebel, and Nedelsky methods, confirm that different methods produce different standards (Brennan & Lockwood, 1980; Kleinke, 1980; Skakun & Kling, 1980).

The discrepancies between standards are large enough to cause important differences in the percentage of students who pass the test. In Skakun and Kling (1980), for example, the Nedelsky cutoff score failed 23% of the General Surgery candidates and a variant of the Ebel method failed 46%. Hambleton (1978) countered that these disparities do not indicate the invalidity of the methods because they reflect predictable differences in how the methods define mastery. But, in Shepard (1980), it was argued that this is not an acceptable defense because the differences in definition are not explicit and are not available to the user. Hambleton (1978) and other measurement experts can see from the standard-setting algorithms that different approaches imply different definitions of minimal competence; but the classroom teacher or school board member seeking to set standards has no guidelines for seeing how different ways of verbally defining the construct of competence can be linked to different quantification strategies. In measurement, different results are acceptable if the intent was to measure different things; but when the labels of mastery or minimal competence from different methods are used interchangeably, then congruence is necessary for validity. If the methods are aimed at different constructs, the labels should be modified to characterize those differences. The discussion of the Nedelsky (1954) method in this paper and that provided by Brennan and Lockwood (1980) are the first efforts to explain how differences in quantification rules imply differences in conceptualization.

A more fundamental problem for judgmental methods is the disagreement among individual judges, even when the same approach is used. Andrew and Hecht (1976) reported general agreement in the passing score set by two groups of judges; Bernknopf, Curry, and Bashaw (1979) also found consensus in the averages obtained from panels of judges. Agreement in averages, however, obscures the sometimes profound variability in individual standards. When panels of judges are randomly equivalent, their means will naturally differ only by sampling error. But consistent averages are not

evidence of consensus; in studies where variability of judges within panels has been reported, the differences have been considerable (Brennan & Lockwood, 1980; Koffler, 1980; Skakun & Kling, 1980).

Three different kinds of advice can be offered for coping with the threat to validity implied by extreme ranges in judges' standards. First, ensure that different value positions and areas of expertise are systematically represented when judges are empanelled. This is more important for high school competency programs than for classroom tests or professional certification examinations because the issues are political and there are more relevant audiences who hold stakes in the testing process (see Jaeger, 1978; Shepard, 1976, 1979). Second, collect evidence of important differences of opinion and consider their significance for validity rather than hiding variability with the group average. Brennan and Lockwood (1980) suggested a reconciliation precedure to have judges meet and arrive at a final standard. In this way, the reason for choosing a particular version of the standard is more explicit. In Kleinke (1980), the judge who specialized in a particular area had the greatest weight; for the North Carolina high school competency test, Jaeger (1978) advised that the final standard be the lowest one set by groups representing different constituencies.

The final recommendation for dealing with subjective and varied standards is to collect validity evidence. Several of the major categories of method that follow can be thought of as alternative strategies for providing different insights to the standards problem *or* as validation strategies to reflect on how sensible the cutoff scores are that have been based on a logical study of the test.

Standards Based on Judgments about Groups

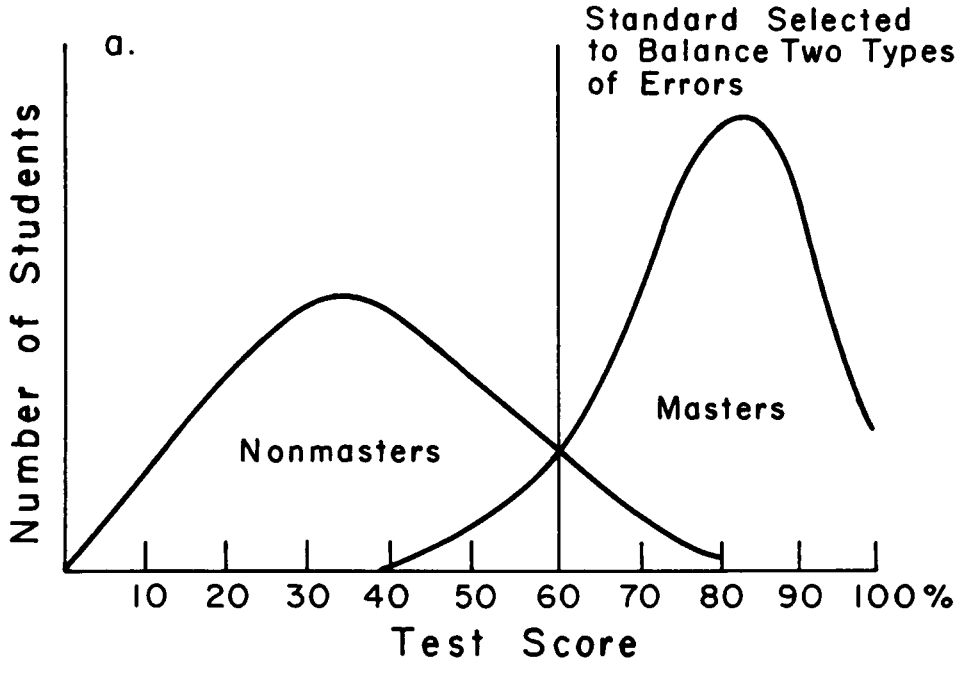
Judgments based on test content are not only inconsistent from one judge to another, they are sometimes obviously wrong. This happens when many individuals believed to be competent fail the test or, conversely, if no one failed the test when it was known that there were nonmasters present. Schoon, Gullion, and Ferrara (1979) noted the tendency of experts to set minimum levels that are unrealistically high:

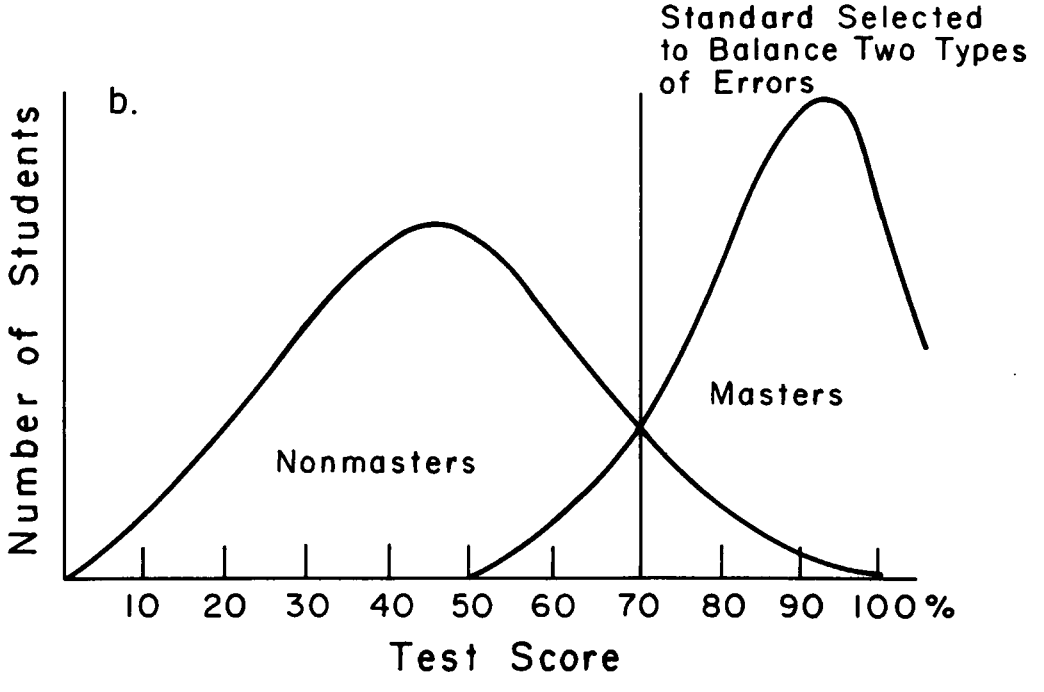
> In our experience with expert committees which set minimum criterion levels, using the (Ebel and Nedelsky) methodologies presented herein, levels are often set that would have failed more than half the candidates. These candidates have all completed accredited educational programs and field work experience under close supervision, and our subjective prior probabilities are that the great majority of these candidates are competent to practice at the entry level. (p. 199)

Sometimes the other evidence of mastery is more compelling than the belief in the validity of the standard.

To avoid the problems of standards that do not agree with the recognition of competence in individuals, standards can be based directly on judgments about the performance of mastery and nonmastery groups. The Contrasting Groups method was proposed by Zieky and Livingston (1977). Teachers or judges are asked to identify individuals who are clearly masters and clearly nonmasters (using information apart from the test); then, the test score distributions for the two separate groups are examined to select the cutting score that best distinguishes them. Figure 1a illustrates a standard set where the curves cross; this choice minimizes the overall classification errors. As will be discussed in a later section, the authors also allow for the cutoff score to be moved up or down to reduce selectively false positive or false negative errors. Koffler (1980) described a more sophisticated statistical approach using a quadratic discriminant function for selecting the point that best discriminates between the groups, but the rationale for the method is still the same.

Figure 1
Changes in the Contrasting Groups Standard Caused by Differences in the Stringency of Nonmastery Classifications





This procedure for finding a dividing score between mastery and nonmastery groups is analogous to Known Groups validation. Just as the validity of a personality instrument is enhanced when it discriminates between clinically identified populations, so the Contrasting Groups method is intended to select the cutoff score which best separates the criterion groups.

Although the additional validity evidence of this method adds an important perspective not provided by judgments of test content, the Contrasting Groups method does not avoid the subjectivity and arbitrariness of standard setting. The simplicity and precision of Figure 1a is misleading. The point of overlap between the curves can vary tremendously, depending upon the judges' definitions of mastery. For example, if judges are stringent and mentally classify marginal students as nonmasters, the standard would drift upward because of a greater range of test scores for the nonmastery group. Figure 1b illustrates how the standard for Figure 1a would change if uncertain cases tended to be sorted as nonmasters rather than being cast into the two groups equally. Even if the instructions to judges improve the certainty of the classifications by suggesting that marginal cases be discarded, the standard will still shift, depending upon different conceptualizations of marginal performance and whether they are deleted equally from the two groups.

The practical importance of judges' conceptualizations of mastery is exemplified in the study by Koffler (1980). Using the Constrasting Groups method, a standard was set for the 11th-grade mathematics competency test given in New Jersey; because there was so much overlap in the distributions for teacher-identified groups of masters and nonmasters, the statistically selected standard passed 100% of the examinees. As Koffler notes, this is clearly unacceptable for a competency testing program chartered to weed out incompetents. One of the possible explanations for why this anomaly arose specifically at the 11th grade level (tests were also given in Grades 3, 6, and 9) is that high school dropouts had removed the worst instances of nonmastery from the judgmental process. Hence, those classified as nonmasters were those who would have been considered marginal in a different context.

The Borderline Group method (Zieky & Livingston, 1977) is also based on judgments of groups. From the same judgmental process described above, only the questionable or borderline cases are obtained and the standard is set at the median of this group's test scores. However, since it is much more difficult to obtain an adequate sample of only borderline examinees, this approach has nothing to recommend it over the Contrasting Groups method.

The Use of Norms

It is a short step from using validation groups to set standards, to using norms to influence the selection of a standard. Relying on normative comparisons to choose a cutoff score initially seems contradictory to the purpose of criterion-referenced testing. After all, criterion referencing was introduced to describe better what an examinee actually knows rather than his/her relative standing in a group. Upon reflection, however, it is only the first use of criterion-referenced testing, estimating domain scores, that can be accomplished without relative comparisons. Qualitative judgments about the excellence or adequacy of performance depend implicity on how others did on the test. Expectations about what a lawyer or high-school graduate should know are normative. If everyone could intuit the theory of relativity on their way to work, Einstein would not have been considered a genius. Similarly, what is now considered "minimal" for a physician to know is based not only on the state of the art, but also on what other professionals have been able to master.

Two points can be made to support the use of normative data in arriving at a cutoff score: (1) this procedure is close to the judgments-about-groups approach, and (2) it provides for direct consideration of acceptable passing rates. Hambleton, Powell, and Eignor (1979) supported other authors in ar-

guing for the supplemental use of norms in standard setting (see also Conaway, 1979; Jaeger, 1978; Shepard, 1976, 1979, 1980), but they disagreed with Glass (1978b) that the 50th percentile of graduating seniors was an appropriate passing score for the California High School Proficiency Examination, a test which allows 16-year-olds to leave high school early. "What," they ask, "can be said of a procedure where whether or not an individual passes or fails. . . depends upon the other individuals taking the test?" (p. 16). The first answer is that the standard was based on a very relevant comparison group. The percentile was obtained from a representative sample of high school seniors and would not change quixotically with the abilities of examinees taking the test on a given date. Secondly, if one considers the true fuzziness in finding the cutting point between mastery and nonmastery groups, the Contrasting Groups method is not fundamentally different from deciding logically on a *range* in the percentile distribution which represents the transition from mastery to nonmastery.

The justification for directly choosing a normatively determined cutoff is best given by the following conjecture. Suppose the California passing score had been determined absolutely by making judgments of test content? What if, in subsequent validation studies, this standard was found to correspond to the 75th percentile of high school seniors? Such a standard would have been unacceptably high; it would have been perceived as subverting the legislative intent for the test which was to allow those who know the *essentials* to leave high school early. Conversely, if the standard fell at the 15th percentile, it would have been thought too generous and would have provoked fears that incompetents were being let out without sufficient schooling. Why not address directly the question of what percentile rank, in the distribution of a relevant and representative criterion group, best corresponds to the judges' conception of mastery? When judges are knowledgeable about the typical range of performance, this is effectively the same as asking them to implement the Contrasting Groups method.

Norms have also been eschewed for criterion-referenced tests because of the belief that placement and certification decisions should be quota-free. Although it is true that for purposes of individualizing instruction the teacher will be more interested in an absolute judgment about whether a student knows enough to benefit from instruction on a subsequent topic, even in the classroom there are relative quota-based decisions to be made. For example, a student's placement in a reading group will depend not only on what he/she knows but also on the need to keep the groups manageable in size and relatively homogeneous in skill level. The student "just between" two performance levels will be shifted to accommodate these practical requirements. For competency and certification tests outside the classroom, there will be implicit quotas created both by practical considerations and the validity issues discussed previously.

Glass (1978b) argued that since any differences in passing rates would be attributable to arbitrariness in the standard-setting methods, standards should be set directly by deciding on an acceptable proportion to fail. There will often be market-place contingencies that govern passing rates. Millman (1973), for example, suggested that the financial cost of providing remedial instruction be taken into account in adjusting the standard. Ignoring political and practical effects of different passing rates can be embarrassing. Rentz (1980) recounted the story of the Georgia teacher certification examination administered during a period of teacher shortage. The test was developed to reflect only minimal content; a passing score was set to indicate the very least one could know on the test, then that standard was adjusted downwards by three standard errors of measurement just to be safe. When the test was finally administered, too few teachers passed to meet the demand, so the standard was thrown out and a new one established more consistent with the desired number of qualified candidates. Kleinke (1980) noted that the examining committee responsible for the National Licensing Examination in Landscape Architecture specifically switched from the Angoff (1971) to the Nedelsky (1954) approach to produce higher passing rates consistent with state boards' expectations. These examples

may horrify those who believe in absolute standards. Absolute standards allow everyone to pass if they are all competent and no one to pass if they are all incompetent. These extremes are rarely encountered. Usually, one finds that there is a large range of credible absolute standards where the distinction between mastery and nonmastery is cloudy. Within the range where competence is uncertain, the sensible way to set a standard is by directly addressing the issue of passing rates.

Similar reasoning for explicitly considering credible passing rates is reflected in Hofstee's (1980) compromise model for establishing cutoff points in which judges are asked to specify the following values:

> First, the maximum required percentage of mastery k_{max} is established. This may be defined as the cutoff score which would be satisfactory even if every student would attain that score at the first trial.
> Second, the minimum acceptable percentage of mastery k_{min} is determined. This level may be defined as the cutoff score below which one would not go even if no student would attain that score at the first trial.
> Third, the maximum acceptable percentage of failures f_{max} is established.
> Fourth, the minimum acceptable percentage of failures f_{min} is established. In our student-centered times, it may seem obvious to set this percentage at zero, but one might argue that this solution is unrealistic, since the next cohort of students will quickly adapt to such a lenient state of affairs and will turn it into a self-defeating policy. (p. 13)

Hofstee, then, uses a graph of the two dimensions—test score and percent passing—and a specific formula for arriving at a midpoint between f_{min}, k_{max} and f_{max}, k_{min}. The result is a compromise between absolute and relative standards.

Empirical Methods for Discovering Standards

The Contrasting Groups approach might have been called an empirical method, since it involves collecting actual data on test performance. Its judgmental aspect was stressed, however, because this is its more salient feature that makes it more useful than straight empirical methods. The empirical methods discussed in this section were intended to avoid subjective judgments. The result is either that the inherent judgments are hidden or that the model of educational outcomes has limited applicability.

Berk (1976) proposed an empirical method nearly identical to the Contrasting Groups method. However, he sought to eliminate the problem of defining (and judging) mastery and nonmastery by selecting two criterion groups: one instructed and the other uninstructed. Some subjectivity is required, since the instructed group must have received "effective" instruction before one can presume that they are masters. The cutoff score that maximizes the agreement between the test classifications and criterion groups is selected.

Berk's original intent was to use this approach with short criterion-referenced tests in instructional settings. Hambleton and Eignor (1979) concluded that the method is promising for this purpose. However, even in the situations for which the method is most appropriate, there will not be a "true" standard to be discovered. The optimal cutoff score identified will depend upon the degree of nonmastery in the uninstructed group and upon both the duration and effectiveness of the instruction received by the group expected to be masters. For minimum competency testing, the method is not applicable at all because it is impossible to identify instructed and uninstructed groups for the competencies tested, since the skills are presumably acquired during 12 years of schooling. Moreover, the

presumption that an instructed group will be predominantly masters is hardly valid; obviously, if instruction guaranteed mastery, the need for minimum competency testing programs would not have arisen in the first place.

Block (1972b) developed a standard-setting model called "educational consequences," named for its attempt to maximize subsequent learning or some other valued outcome. The method depends on there being a functional relationship between performance on the test and level of attainment on the criterion variable. The curve is expected to look like a learning curve, as illustrated in Figures 2a–2d. Experimental studies must be carried out using different mastery cutoffs to determine the nature of the relationship; then, the cutting score on the test (C) is selected to maximize performance on the outcome dimension.

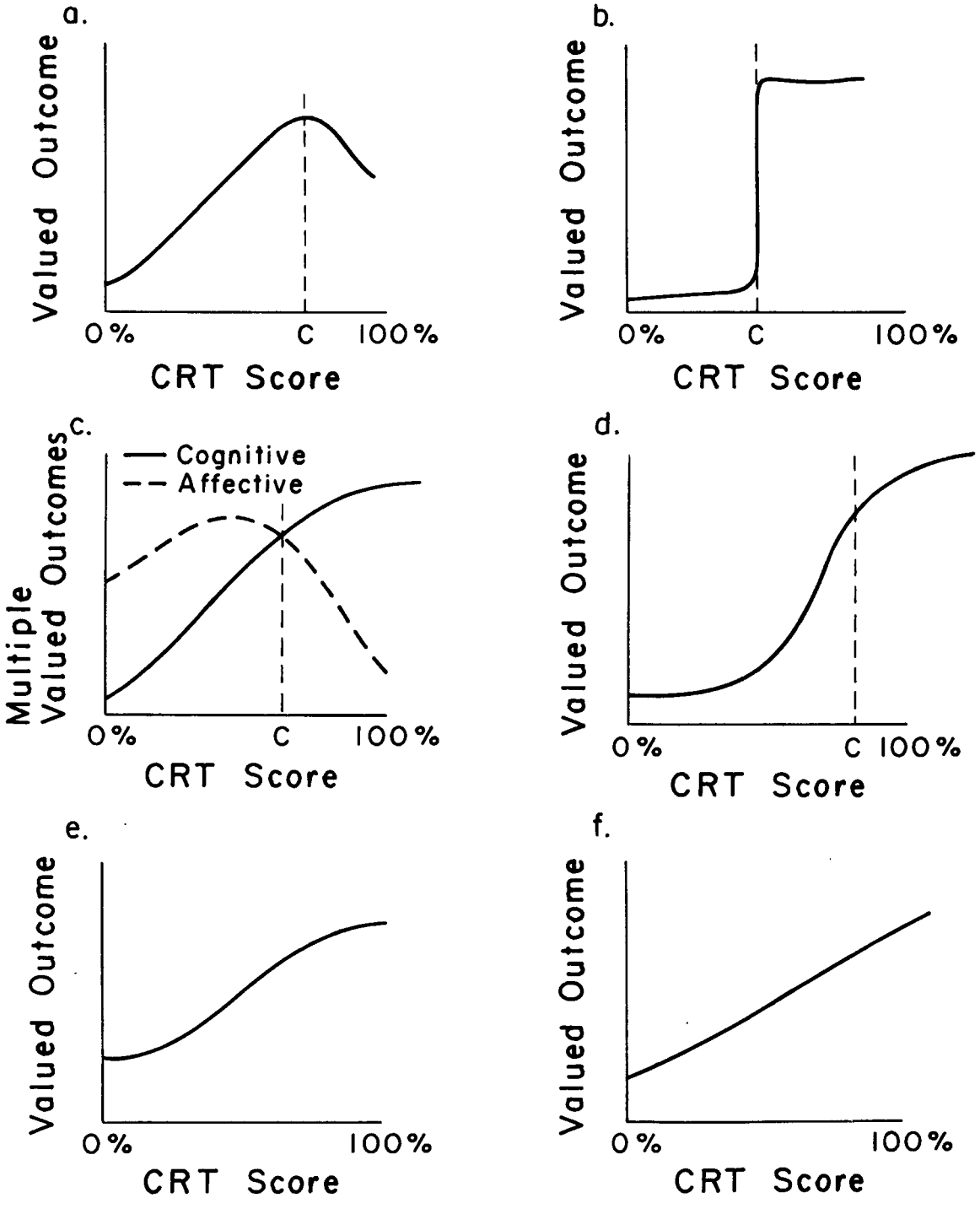
Glass (1978b) severely criticized this method, which was intended to discover an appropriate standard scientifically. Unless the relationship between the test and the valued outcome is nonmonotonic (i.e., increases and then decreases), a 100% test standard will be optimal. This is obviously unrealistic and hardly worth the extensive investment in field trials to determine. Figures 2a and 2b illustrate the kinds of functions that would have to occur to make the desired standard obvious. Curve 2a is nonmonotonic, indicating that increasing one's score beyond a certain point on the criterion-referenced test actually diminishes performance on the subsequent task or criterion variable; it is difficult to imagine a cognitive task for which this relationship would occur. Curve 2b is a step function; although it is more plausible than 2a, it will rarely occur in practice. Curves 2d and 2e are more gradual and better resemble real data (in Block, 1972b, actual shapes were flatter than 2e). These more realistic relationships require judgments to arrive at a cutoff point; someone has to decide how much of the criterion variable is good enough. Block hoped to escape this ambiguity by using multiple valued outcomes.

If an attitude variable with a decreasing relationship to test performance could be found, then the composite criterion would create the desired nonmonotonic curve, as shown in 2c. However, even if such a graph were obtained, it is still a matter of choice whether the outcome variables should be weighted equally in arriving at a composite; the standard will shift as the weights shift. Finally, it should be noted that if performance on the test and the subsequent task are not hierarchically related, there may not be any interesting or informative bends in the curve at all, as in Figure 2f. Given the very limited success of research intended to demonstrate the existence of learning hierarchies, it is unlikely that this approach will solve the standard-setting dilemma even in delimited classroom subjects. The method is even less likely to work for minimum competency testing programs where it is impossible to reach agreement about workable operationalizations of the criterion variable. As Hambleton and Eignor (1979) stated, "One can't maximize a valued outcome if the outcome can't be defined in any reasonable manner."

Empirical Methods for Adjusting Standards

By far, the largest number of standard-setting methods fall into this last category. They include the more technical methods and convey the impression that standards can be determined with scientific precision. In truth, however, these approaches do not determine a standard; rather, they presume that a standard already exists on an external criterion and merely translate this into a cutoff score on the test. For classroom purposes, the criterion variable might be success on a later learning assignment; for medical certification, the criterion might be performance tests in actual clinical practice. Since the criterion dichotomy is not an all-or-none state but rather an arbitrary break on a continuum, this means that someone has already had to wrestle with the standard-setting choices.

Figure 2
Hypothetical Relationships Between a Criterion-Referenced Test Score and a Valued Outcome



These methods for locating the passing score on the test are based on decision theory (see Cronbach & Gleser, 1965; Hambleton & Novick, 1973). The object is to match the test dichotomy to the criterion dichotomy to ensure the smallest number of classification errors. The familiar four-fold table, given below, represents the two kinds of correct decisions and two kinds of incorrect decisions that can be made in any placement or certification situation:

		External Criterion: Nonmaster $\pi < \pi_o$	External Criterion: Master $\pi \geq \pi_o$
Test-Based Decision: Fail $X < C$		*Correct Nonmasters*	*False Negatives*
Test-Based Decision: Pass $X \geq C$		*False Positives*	*Correct Masters*

This conceptualization is the same as that offered by other authors as a model for decision validity; given the standard on the criterion, the cutoff score on the test is selected to maximize validity. The approach is not very different from the Contrasting Groups method, except that the validation variable must have a specific demarcation point instead of two overlapping groups.

Huynh's (1976) empirical Bayesian approach is one of the better known procedures for setting cutoff scores, given the existence of an external criterion. Huynh was initially interested in situations where criterion-referenced tests would be used to determine when a student's mastery of a topic was sufficient to allow him/her to progress to the next topic. Success on the next unit of instruction, called the referral task, is used as the criterion. The reasoning is the same as for Block's educational consequences method, except that there is already somehow a standard for success on the outcome variable. Given π_o, the cutoff score for success on the criterion task for the test (C) is chosen, so that in the table above the average loss, $P(\pi < \pi_o, X \geq C) + P(\pi \geq \pi_o, X < C)$, is the smallest. Other methods following the same paradigm are those by Davis and Diamond (1974), Kriewall (1969, 1972) and Livingston (1976). Glass (1978b) called these methods "bootstrapping on other criterion scores" and faulted the authors for taking at face value the standards that must already exist, e.g., the very frequent 80% on criterion-referenced tests and the traditional 70% on competency tests. Hambleton and Eignor (1979) also found little to recommend these methods for minimum competency testing programs, since no agreed upon criterion measure or standard of adult success is likely to be found.

Several of the decision-theoretic methods have an additional feature worthy of note. These authors and methods have given more attention than most to weighing the differential costs of two types of classification errors. Originally, Millman (1973) recommended techniques for adjusting a cutoff score to protect against the more serious type of error. For example, in an instructional context, the possibility of failing on the next learning unit has to be balanced against the boredom and annoyance of repeating a task that has already been mastered. In minimum competency testing programs, the consequences of decision errors are more serious. There are different opinions about whether standards should be set high but should give three chances, or should be set low to prevent unfair failures.

Various mathematical models are available for reducing either false-positive or false-negative errors on the basis of assigned utilities. Procedures proposed by Livingston (1975) and van der Linden and Mellenbergh (1977) use linear functions to quantify the expected loss from the two kinds of error.

The loss function to be minimized is like Huynh's equation to minimize the proportion of incorrect classifications, but now weights have been assigned according to the type and extent of the loss. The Bayesian method introduced by Novick and Lewis (1974) adds the use of prior information on examinees as well as specifying utilities in the form of loss ratios. Although this approach is quite complete, it is difficult to know when it will be useful practically. In classroom situations, it will probably be too cumbersome. In minimum competency testing applications, it would probably be worth the extra complexity and cost; but it is unlikely that prior data on examinees could be introduced without raising more validity issues than the test itself. Novick and Lindley (1978) gave further attention to the appropriate form for the utility function; the normal distribution and other families of distributions are probably an improvement on simple linear functions for representing gains and losses associated with decision errors.

All of these methods still presume that the standard-setting problem has already been solved for the criterion variable. They also presume that the teacher or administrator will know how to assign the necessary utilities (and to choose the right shape for the utility function). Virtually no advice is given about how to try and quantify both financial and psychological costs and benefits, e.g., the job-getting potential of a high school diploma for a marginal student versus the cost to society when the diploma is devalued by unqualified graduates. Probably these discussions have been avoided because there is very little that can be offered concretely except to say that one has to "sort of" choose numbers that roughly reflect one's values. The use of loss ratios, "one type of mistake is twice as bad as the other," probably better conveys the global and subjective nature of the judgments. Still, decision makers will have a difficult time arriving at specific numbers. They will know that they consider false positives to be more serious than false negatives but not whether they are half-again as costly or three times more costly. Because such numbers are highly subjective, administrators are urged to consider how changes in credible loss ratios, 2-to-1 or 3-to-1, will affect the passing rates.

SELECTION OF STANDARD-SETTING METHODS FOR SPECIFIC USES

Pupil Diagnosis

Classroom passing scores for instructional placement are frequently set informally because classroom teachers do not have the resources for elaborate standard-setting methods. Furthermore, teachers can tolerate more errors in judgment because misclassifications can easily be detected and corrected. If a nonmaster is accidently moved on to the next topic because of a low standard, his/her incomplete knowledge will be noticeable in struggles with the new material. If the material is not hierarchical, the next review test should catch the error.

The best advice for the teacher is to keep in mind both absolute and normative conceptualizations of mastery. For example, what are the expectations for a passing essay about the causes of the Civil War? and what are typical experiences with essays written by eighth graders considered to be masters? In formal terms, this means reconciling the insights provided by judgments about test content and judgments about groups.

When subject matter is sequential, teachers could also benefit from an understanding of the empirical validation methods. What Block (1972a) did with experimental variation of the mastery score, teachers can do by trial and error. If there is a noticeable number of apparent nonmasters struggling on a new unit, the previous standard may have been too low. Other possible explanations for widespread learning difficulties are poor instructional methods or students missing other prerequisite skills not measured on the previous test. If these possibilities are considered and ruled out, then it may be advisable to adjust the standards to see if more reasonable placements result. What is being

sought is a standard similar to Huynh's that will ensure success on the referral task. There are always some classification errors; but if the teacher's impression is that too many are being held back or too many passed on with incomplete mastery, then the passing score should be moved.

For objectives-based instructional programs implemented for an entire school district, it would be possible to set standards formally and to collect validity data systematically. Meanwhile, researchers should continue to establish the existence of learning hierarchies and to discover whether perservering on the first task will improve success on the second. With or without verifiable hierarchies a simpler question to be addressed by research is, what level of mastery assures long-term retention? No matter how fruitful these inquiries are, it is certain that the optimal passing score will be different for each test and for each learning context. The only generalizable findings are likely to be procedures for gathering validity data and adjusting cutoff scores accordingly.

Standard-setting methods may cause teachers to think more about their decisions to review or go on to the next topic, but they will not provide a new science to solve the dilemma of when to push ahead despite incomplete knowledge.

Pupil Certification

Certification standards for high school graduation or professional licensing require a composite approach to protect against the fallibility of separate methods. This strategy is analagous to using multiple measures for triangulation when operationalizations of research outcomes are imperfect (Webb, Campbell, Schwartz, & Sechrest, 1966).

At a minimum, standard-setting procedures should include a balancing of absolute judgments and direct attention to passing rates. All of the embarrassments of faulty standards that have ever been cited are attributable to ignoring one or the other of these two sources of information. If absolute judgments are ignored, incompetent doctors could pass the test if they were members of a weak class. High school seniors are sometimes graduated without basic skills because this is the norm. Since criterion-referenced testing was developed to overcome the problems of relative judgments, this error is not usually made with criterion-referenced tests. Instead, out of loyalty to absolute standards, examining boards have made the opposite error of setting standards without norms that fail half the medical school class or that fail to fail any high school graduates in an entire state. Direct attention to passing rates will allow standard setters to reconcile their beliefs about the required competencies (items on the test) and their beliefs about how many individuals are qualified.

The Angoff (1971) and Jaeger (1978) methods for judging test content are recommended as the most practical. As part of their deliberations, judges should have normative data to consider (Conaway, 1979; Hambleton et al., 1979; Jaeger, 1978; Shepard, 1976, 1979; Zieky & Livingston, 1977). In addition, Shepard (1980) proposed that judges make independent estimates of anticipated failure rates.

When the standard setters are teaching experts for a particular profession, they may have a good sense of how students they know will perform on the test. When judges are removed from the population of test takers, as they may be when political constituencies set the standards for high school competency tests, additional validity data may be needed to confirm the wisdom of the standard. The Contrasting Groups method (Zieky & Livingston, 1977) is the preferred method for verifying whether the dichotomy on the test corresponds to the distinction between individuals who are judged to be masters and nonmasters. Beyond this, however, for certification purposes, the empirical methods for discovering standards (Berk, 1976; Block, 1972b) add nothing to the method based on judgments about groups because no valued outcome or criterion dimension can be operationalized. Similarly,

the various statistical techniques that presume the existence of a standard on an external criterion (Huynh, 1976; Livingston, 1975, 1976; van der Linden & Mellenbergh, 1977) are not applicable. The methods that propose utility functions to quantify loss ratios (Novick & Lewis, 1974; Novick & Lindley, 1978) may be helpful in conceptualizing adjustments in the cutoff score to compensate for differences in the costs of classification errors. From all of these methods, the best composite approach will reconcile absolute judgments with empirical data with some additional consideration for weighing the two types of error.

Program Evaluation

Standards are subjective and variable. They are arbitrary cutting points along a continuous performance scale. Because standards impose an artificial dichotomy, they obscure performance information about individuals along the full performance continuum. Therefore, standards should not be used to interpret test data regarding the worth of educational programs. For other uses of criterion-referenced tests, it could be argued that pass/no-pass results were essential to serve educational decisions; but for program evaluation purposes, there are other more appropriate ways to attach value to the goodness or badness of the test results. Dichotomous classifications of individuals are not needed.

Program evaluation interpretations tend to compound the errors in separate standards. For example, if 10% of the students fail the reading test and 35% fail the math test, there is no way to tell whether this discrepancy is due to better teaching in reading or to a more lenient standard (see Glass's, 1978a, criticism of the Florida Functional Literacy Test interpretations). Variability in the stringency of the standards will be mistaken for program strengths and weaknesses. When standards are further layered to require that 80% of the students attain the mastery criterion on at least 80% of the objectives, less and less light is shed on whether the program is better than other programs or if the students are learning as much as they should be.

When group achievement is only reported as a percent who passed the standard, the reader has no sense of whether this is an unusually good or bad rate. Popham (1976) suggested that normative data be added to criterion-referenced test results to supply comparative meaning. But comparisons can better be made with means or even quartile scores rather than with percent passing. Using passing rate as a group statistic means that achievement gains will only be reflected if they occur near the cutoff score. The effects of the program for those already above the standard are ignored. For some time evaluators have been concerned that gains in the group mean could occur because of growth in only one subgroup. To ensure that the program is effective for the full range of students, gains can be mapped at quartile points as well as the mean. With percent passing as an index of program quality, there is no way to compensate for its insensitivity to changes in performance that are not near the cutoff score.

FUTURE RESEARCH ON STANDARD SETTING

Recommendations for future research are offered with some reluctance because such a call may seem to imply that redoubled efforts will eventually produce a nonarbitrary methodology. This hope, however, is more like an alchemist's belief than the reasonable confidence of a scientist seeking a cure for cancer. Earlier discussions were meant to emphasize that it is the nature of the problem and not the immaturity of the technology that leads to artificial dichotomies with poor validity near the cutting point. This is a permanent dilemma for standard setters.

Since it is not possible to discover an objective, nonjudgmental method, research aimed at improving standard setting should be focused on practical and procedural questions such as the following:

What are the effects of different instructions to judges?
What are the advantages and disadvantages of having judges work independently or in groups?
What format is most useful for presenting normative data to judges? At what stage in the process should these data be considered?
When absolute standards are contradicted by validity data, what factors influence the final choice of a standard?
What strategies do decision makers find useful in assigning specific utilities to classification errors?

Applied studies of this kind are probably only warranted for certification uses of criterion-referenced tests. Then the testing is on a large enough scale and the consequences of the standards are important enough to merit formal research and development.

There are also larger issues that should ultimately be addressed by evaluation studies or basic research. Criterion-referenced testing was developed to improve instruction and learning. The effectiveness of objectives-based instructional systems can be compared to more traditional models of teaching and testing. In this context, systematic study of the effect of different cutoff scores on long-term retention or on successful transfer to a subsequent task will greatly inform judgments about standards even if it does not magically produce a cutoff score. At the same time, if different cutoff scores can be shown to make important differences on other learning tasks, valuable evidence is gained to support the sequential learning model itself. The effects of having a test and enforcing a standard can also be evaluated for certification tests such as minimum competency tests for high school graduation. The National Institute of Education has recently commissioned an evaluation of minimum competency testing programs in the United States, which should address such questions as whether the presence of standards improves the knowledge of marginally passing students beyond what it would be under traditional graduation requirements. That is, without the incentive of the test, would such students have scores in the incompetence range? For both classroom placement and pupil certification uses of tests, for which pass-fail decisions are essential, the positive consequences of having admittedly arbitrary standards are believed to outweigh the costs. This is a researchable question.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. *Standards for educational and psychological tests.* Washington, DC: American Psychological Association, 1974.

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement,* 1976, *36*, 35-50.

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

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education,* 1976, *45*, 4-9.

Bernknopf, S., Curray, A., & Bashaw, W. L. *A defensible model for determining a minimal cut-off score for criterion-referenced tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979.

Block, J. H. *Student evaluation: Toward the setting of mastery performance standards.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972. (a)

Block, J. H. Student learning and the setting of mastery performance standards. *Educational Horizons,* 1972, *50*, 183-190. (b)

Block, J. H. Standards and criteria: A response. *Journal of Educational Measurement*, 1978, *15*, 291-295.

Brennan, R. L., & Lockwood, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, 1980, *4*, 219-240.

Conaway, L. E. Setting standards in competency-based education: Some current practices and concerns. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency based education.* Washington, DC: National Council on Measurement in Education, 1979.

Cronbach, L. J. *Essentials of psychological testing* (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1971.

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

Davis, F. B., & Diamond, J. J. The preparation of criterion-referenced tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

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

Educational Testing Service. *Report on a study of the use of the National Teachers Examination by the state of South Carolina.* Princeton, NJ: Educational Testing Service, 1976.

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

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

Glass, G. V. Minimum competence and incompetence in Florida. *Phi Delta Kappan*, 1978, *59*, 602-605. (a)

Glass, G. V. Standards and criteria. *Journal of Educational Measurement*, 1978, *15*, 237-261. (b)

Hambleton, R. K. On the use of cutoff scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 1978, *15*, 277-290.

Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard setting. In R. Jaeger & C. Tittle (Eds.), *Minimum competency testing.* Berkeley, CA: McCutchan, 1979.

Hambleton, R. K., Powell, S., & Eignor, D. R. Issues and methods for standard setting. In R. K. Hambleton & D. R. Eignor, *A practitioner's guide to criterion-referenced test development, validation, and test score usage* (Laboratory of Psychometric and Evaluative Research Report No. 70, 2nd ed.). Amherst: University of Massachusetts, School of Education, 1979.

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*, 1978, *48*, 1-47.

Hofstee, W. K. B. *Policies of educational selection and grading: The case for compromise models.* Paper presented at the Fourth International Symposium on Educational Testing, Antwerp, Belgium, June 1980.

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

Jaeger, R. *Measurement consequences of selected standard-setting models.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Jaeger, R. M. *A proposal for setting a standard on the North Carolina High School Competency Test.* Paper presented at the spring meeting of the North Carolina Association for Research in Education, Chapel Hill, 1978.

Jaeger, R. M. Measurement consequences of selected standard-setting models. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement.* Washington, DC: National Council on Measurement in Education, 1979.

Kleinke, D. J. *Applying the Angoff and Nedelsky techniques to the National Licensing Examinations in Landscape Architecture.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Koffler, S. L. A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement*, 1980, *17*, 167-178.

Kriewall, T. E. *Application of information theory and acceptance sampling principles to the management of mathematics instruction.* Unpublished doctoral dissertation, University of Wisconsin, 1969.

Kriewall, T. E. *Aspects and applications of criterion-referenced tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

Linn, R. L. Issues of validity in measurement for competency-based programs. In M. A. Bunda & J. R. Saunders (Eds.), *Practices and problems in competency-based measurement*. Washington, DC: National Council on Measurement in Education, 1979.

Livingston, S. A. *A utility-based approach to the evaluation of pass/fail testing decision procedures* (Report No. COPA-75-01). Princeton, NJ: Educational Testing Service, 1975.

Livingston, S. A. *Choosing minimum passing scores by stochastic approximation techniques* (Report No. COPA-76-02). Princeton, NJ: Educational Testing Service, 1976.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research,* 1976, *45*, 133-158.

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

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications.* Berkeley, CA: McCutchan, 1974.

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

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurements. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Novick, M. R., & Lindley, D. V. The use of more realistic utility functions in educational applications. *Journal of Educational Measurement,* 1978, *15*, 181-191.

Popham, W. J. *Educational evaluation.* Englewood Cliffs, NJ: Prentice-Hall, 1975.

Popham, W. J. Normative data for criterion-referenced tests? *Phi Delta Kappan,* 1976, *58*, 593-594.

Popham, W. J. As always provocative. *Journal of Educational Measurement,* 1978, *15*, 297-300.

Rentz, R. R. *Discussion*. Presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Roudabush, G. E. *Models for a beginning theory of criterion-referenced tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1974.

Schoon, C. G., Gullion, C. M., & Ferrara, P. Bayesian statistics, credentialing examinations, and the determination of passing points. *Evaluation & The Health Professions,* 1979, *2*, 181-201.

Scriven, M. The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (AERA Monograph 1). Chicago: Rand McNally, 1967.

Scriven, M. How to anchor standards. *Journal of Educational Measurement,* 1978, *15*, 273-275.

Shepard, L. A. Setting standards and living with them. *Florida Journal of Educational Research,* 1976, *18*, 23-32.

Shepard, L. A. Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement.* Washington, DC: National Council on Measurement in Education, 1979.

Shepard, L. A. Technical issues in minimum competency testing. In D. C. Berliner (Ed.), *Review of research in education* (Vol. 8). Itasca, IL: F. E. Peacock Publishers, in press.

Skakun, E. N., & Kling, S. Comparability of methods for setting standards. *Journal of Educational Measurement,* 1980, *17*, 229-235.

van der Linden, W. J., & Mellenbergh, G. J. Optimal cutting scores using a linear loss function. *Applied Psychological Measurement,* 1977, *1*, 593-599.

Webb, E. J., Cambell, D. T., Schwartz, R. D., & Sechrest, L. *Unobstrusive measures: Nonreactive research in the social sciences.* Chicago: Rand McNally, 1966.

Zieky, M. J., & Livingston, S. A. *Manual for setting standards on the Basic Skills Assessment Tests.* Princeton, NJ: Educational Testing Service, 1977.

Author's Address

Lorrie Shepard, Laboratory of Educational Research, University of Colorado, Boulder CO 80309.