DOCUMENT RESUME

ED 206 722 TM 810 631

AUTHOR Hambleton, Ronald K.; And Others
TITLE Issues and Methods for Standard-Setting.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 79
NOTE 80p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Academic Standards; *Criterion Referenced Tests; *Cutting Scores; Elementary Secondary Education; *Methods; *Models
IDENTIFIERS *Continuum Models; Empirical Methods; Judgmental Processes; Test Use

ABSTRACT

Issues involved in standard setting along with methods for standard setting are reviewed, with specific reference to their relevance for criterion referenced testing. Definitions are given of continuum and state models, and traditional and normative standard setting procedures. Since continuum models are considered more appropriate for criterion referenced testing purposes, they are examined in greater depth. The continuum models are subdivided into three categories: judgmental methods; empirical methods; combinations of judgment and empirical methods. For the purpose of viewing a test as an entity and not in relation to other variables, the Angoff and Nedelsky judgmental methods are considered useful; however, when empirical data is available, the empirical methods delineated in Berk's method of the Contrasting Groups method are recommended as being appropriate. Discussed in the final section of the paper are procedures for setting standards to accomplish three primary uses of criterion referenced testing: classroom testing; basic skills testing for annual promotion and high school graduation; professional licensing and certification testing. However, further research is recommended before the latter procedures are implemented.
(Author/AEF)

# Issues and Methods for Standard-Setting[1,2,3]

*Ronald K. Hambleton, Sally Powell*
*University of Massachusetts, Amherst*

and

*Daniel R. Eignor*
*Educational Testing Service*



[1]Preparation of this material was supported, in part, by a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

[2]Several sections of the material were included in a presentation at an invited symposium on standard-setting at the annual meeting of NCME, San Francisco, 1979.

[3]The material is an up-dated version of Unit 6 from Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, Amherst, 1979. (2nd edition, 480 pages)

## Table of Contents


| | | Page |
|---|---|---|
| 6.0 | Overview of the Unit. . . . . | 1 |
| 6.1 | Introduction. . . . . | 2 |
| 6.2 | Some Issues in Standard Setting . . . . . | 5 |
| | 6.2.1 Uses of Cut-Off Scores in Decision-Making. . . . . | 5 |
| 6.3 | Distinction Between Continuum and State Models. . . . . | 13 |
| 6.4 | Traditional and Normative Procedures. . . . . | 15 |
| 6.5 | Consideration of Several Promising Standard Setting Methods . . . . . | 18 |
| 6.6 | Judgmental Methods. . . . . | 20 |
| | 6.6.1 Item Content . . . . . | 20 |
| | 6.6.2 Guessing and Item Sampling . . . . . | 30 |
| 6.7 | Empirical Methods . . . . . | 31 |
| | 6.7.1 Data From Two Groups . . . . . | 31 |
| | 6.7.2 Decision-Theoretic Procedures. . . . . | 38 |
| | 6.7.3 Empirical Methods Depending Upon a Criterion Measure . . . . . | 43 |
| | 6.7.4 Educational Consequences . . . . . | 48 |
| 6.8 | Combination Methods . . . . . | 52 |
| | 6.8.1 Judgmental-Empirical . . . . . | 52 |
| | 6.8.2 Bayesian Procedures. . . . . | 55 |
| 6.9 | Some Procedural Steps in Standard Setting . . . . . | 57 |
| | 6.9.1 Preliminary Considerations . . . . . | 58 |
| | 6.9.2 Classroom Testing. . . . . | 59 |
| | 6.9.3 Basic Skills Testing for Annual Promotion and High School Graduation. . . . . | 64 |
| | 6.9.4 Professional Licensing/Certification Testing . . . | 68 |
| 6.10 | Summary . . . . . | 70 |
| 6.11 | References. . . . . | 72 |

## 6.0 Overview of the Unit

In this Unit, some of the issues involved in standard setting along with methods for standard-setting are reviewed. The review will draw on the work of Millman, Meskauskas, and Glass and incorporate many of the newer standard-setting methods. The standard-setting methods are organized into three categories, judgmental methods, empirical methods, and combinations of judgment and empirical methods. Procedures for setting standards to accomplish three primary uses of criterion-referenced testing are discussed in a final section of the paper.

## 6.1 Introduction

In a recent review of the criterion-referenced testing field, Hambleton, Swaminathan, Algina, and Coulson (1978) delineated two major uses for test scores derived from criterion-referenced tests: domain score estimation and the allocation of examinees to mastery states. The second use, the allocation of examinees to mastery states, requires the setting of a performance standard, or cut-off score.

Based upon an individual's score on a test, where the test is a representative sample of the subject domain, a mastery/non-mastery decision concerning the domain from which the item sample was drawn is sought. Millman (1973) summarizes the situation well:

> Of interest is the proportion of such items a student can pass. It is assumed that some educational decision, e.g., the nature of subsequent instruction for the student, is conditional upon whether or not he exceeds a proficiency standard when administered a sample of items from the domain. Thus, attention is directed toward the individual examinee and his performance relative to the standard rather than toward producing indicators of group performance.

Thus, it can be seen that in this criterion-referenced testing situation, a cut-off score (there can be multiple cut-off scores on the domain score scale although usually only one is set) must be set, in order to make a decision about an individual's mastery status. The results of this decision will depend upon the context within which the test is being used. As an example, consider the Mastery Learning paradigm (Block, 1972). In this situation, if a student's score exceeds the cutting score, he/she is advanced to the next unit of instruction. If the student's score falls below

the standard, remedial activities are prescribed. It is important to understand that the decision being made is on the level of the individual, and as such, the status of other individuals does not enter into the decision. As a second example, consider the use of criterion-referenced tests to provide test data relative to a set of basic skills which students must demonstrate mastery of (i.e., achieve specified levels of performance) in order to graduate from high school. In this context, decisions are very important because whether or not students can graduate will depend on their criterion-referenced test score performance and the resulting master/non-mastery decisions which are made.

These situations can be contrasted with the setting of standards for norm-referenced tests, which is considerably less complex. Since for tests constructed to yield norm-referenced interpretations, an individual is compared to others, it makes sense to set a passing or cut-off score so that a certain percent of the students pass. If, for instance, only 20% of the students taking an exam can be placed in an enrichment program, then a passing score that passes 20% of the students would make sense.

Given what has just been said about the importance of cut-off scores for proper criterion-referenced test score usage, one would think that this would be well-researched and documented area. This is simply not the case. Most of the work done to date has been concerned with the suggestion of possible methods, perhaps twenty-five in number, rather than with actual empirical investigations. In addition to the individual work done, there have been two excellent

reviews of cut-score procedures advanced (Millman, 1973; Meskauskas, 1976), and one recent review that was highly critical of the field (Glass, 1978a, 1978b).

## 6.2 Some Issues in Standard Setting

One of the primary purposes of criterion-referenced testing is to provide data for decision-making. Sometimes the decisions are made by classroom teachers concerning the monitoring of student progress through a curriculum. On other occasions, promotion, certification and/or graduation decisions are made by school, district, and state administrators.

Glass (1978a) was rather critical of measurement specialists for giving too little attention to the problem of determining cut-off scores [he notes, "A common expression of wishful thinking is to base a grand scheme on a fundamental, unsolved problem." (p. 1)]. On the other hand, a considerable amount of criterion-referenced testing research has been done. Not all uses of criterion-referenced tests require cut-off scores (for example, description), and moreover, the problem does not really arise until a criterion-referenced test has been constructed. Also, it should not be forgotten that problems associated with cut-off scores are difficult and so solutions are going to require more time.

### 6.2.1 Uses of Cut-off Scores in Decision Making

A "cut-off score" is a point on a test score scale that is used to "sort" examinees into two categories which reflect different levels of proficiency relative to a particular objective measured by a test. It is common to assign labels such as "masters" and "non-masters" to examinees assigned to the two categories. It is not unusual either to assign examinees to more than two categories based on their test performance (i.e., sometimes multiple cut-off scores are

used) or to use cut-off scores that vary from one objective to another (this may be done when it is felt that a set of objectives differ in their importance).

It is important at this point to separate three types of standards or cut-off scores. Consider the following statement:

> School district A has set the following target—It desires to have 85% or more of its students in the second grade achieve 90% of the reading objectives at a standard of performance equal to or better than 80%.

Three types of standards are involved in the example:

1. The 80% standard is used to interpret examinee performance or each of the objectives measured by a test.
2. The 90% standard is used to interpret examinee performance across all of the objectives measured by a test.
3. The 85% standard is applied to the performance of second graders on the set of objectives measured by a test.

In this unit, only the first use of standards or cut-off scores will be considered.

In what follows it is important to separate the theoretical arguments for or against the uses of cut-off scores from the uses and misuses of cut-off scores in practical settings. For example, it is well-known that cut-off scores are often "pulled from the air" or set to (say) 80% because that is the value another school district is using. But, the fact that cut-off scores are being determined in a highly inappropriate way is obviously not grounds for rejecting the concept of a "cut-off score." If the concept is appropriate for some particular use of a criterion-referenced test, the task becomes one of training people to set and to use cut-off scores properly (Hambleton, 1978).

Four questions with respect to the use of cut-off scores with criterion-referenced tests require answers:

1. Why are cut-off scores needed?
2. What methods are available for setting cut-off scores?
3. How should a method be selected?
4. What guidelines are available for applying particular methods successfully?

## 1. Why are cut-off scores needed?

An answer to the question depends on the intended use (or uses) of the test score information. Consider first objectives or competency-based programs since it is with these types of programs that criterion-referenced tests and cut-off scores are most often used. Objectives-based programs, in theory are designed to improve the quality of instruction by (1) defining the curricula in terms of objectives, (2) relating instruction and assessment closely to the objectives, (3) making it possible for individualization of instruction, and (4) providing for on-going evaluation. Hard evidence on the success of objectives-based programs (or most new programs) is in short supply but there is some evidence to suggest that when objectives-based programs are implemented fully an d properly they are better than more "traditionally-oriented" curricula (Klausmeier, Rossmiller, & Saily, 1977; Torshen, 1977). Individualization of instruction is "keyed" to descriptive information provided by criterion-referenced tests relative to examinee performance on test items measuring objectives in the curriculum. But descriptive information such as "examinee A has answered correctly

85% of the test items measuring a particular objective" must be evaluated and decisions made based upon that interpretation. Has a student demonstrated a sufficiently high level of performance on an objective to lead to a prediction that she/he has a good chance of success on the next objective in a sequence? Does a student's performance level indicate that he/she may need some remedial work? Is the student's performance level high enough to meet the target for the objective defined by teachers of the curriculum? In order to answer these and many other questions it is necessary to set standards or cut-off scores. How else can decisions be made? Comparative statements about students (for example, Student A performed better than 60% of her classmates) are largely irrelevant. Carefully developed cut-off scores by qualified teams of experts can contribute substantially to the success of an objectives-based program (competency-based program or basic skills program) because cut-off scores provide a basis for effective decision-making.

There has also been criticism (Glass, 1978a) of the use of cut-off scores with "life skills" or "survival skills" tests. The are terms currently popular with State Departments of Education, School Districts, Test Publishers, and the press. Of course, Glass is correct when he notes that it would be next to impossible to validate the classifications of examinees into "mastery states", i.e., those predicted to be "successful" or "unsuccessful" in life. On the other hand, if what is really meant by the term "life skills" (say) is "graduation requirements," then standards of performance for "basic skills" or "high school competency" tests can probably be set by appropriately chosen groups of individuals (Millman, personal communication).

## 2. If cut-off scores are needed, what methods are available for setting them?

Numerous researchers have catalogued many of the available methods (Hambleton & Eignor, 1979; Hambleton et al., 1978; Jaeger, 1976; Millman, 1973; Meskauskas, 1976; Shepard, 1976). Many of these methods have also been reviewed by Glass (1978a). It suffices to say here that there exist methods based on a consideration of (1) item content, (2) guessing and item sampling, (3) empirical data from mastery and non-mastery groups, (4) decision-theoretic procedures, (5) external criterion measures, and (6) educational consequences. These methods will be considered in detail in sections 6.6, 6.7, and 6.8.

What is clear is that all of the methods are arbitrary and this point has been made or implied by everyone whose work we have had an opportunity to read. The point is not disputed by anyone we are aware of. But as Glass (1978a) notes, "arbitrariness is no bogey-man, and one ought not to shrink from a necessary task because it involves arbitrary decisions" (p. 42). Popham (1978) has given an excellent answer to the concern expressed by some researchers about arbitrary standards:

> Unable to avoid reliance on human judgment as the chief ingredient in standard-setting, some individuals have thrown up their hands in dismay and cast aside all efforts to set performance standards as arbitrary, hence unacceptable.
>
> But Webster's Dictionary offers us two definitions of arbitrary. The first of these is positive, describing arbitrary as an adjective reflecting choice or discretion, that is, "determinable by a judge or tribunal." The second definition, pejorative in nature, describes arbitrary as an adjective denoting capriciousness, that is, "selected at random and without reason."

> In my estimate, when people start knocking the standard-setting game as arbitrary, they are clearly employing Webster's second, negatively loaded definition.
>
> But the first definition is more accurately reflective of serious standard-setting efforts. They represent genuine attempts to do a good job in deciding what kinds of standards we ought to employ. That they are judgmental is inescapable. But to malign all judgmental operations as capricious is absurd. (p. 168)

And, in fact, much of what we do is arbitrary in the positive sense of the word. We set fire standards, health standards, environmental standards, highway safety standards, (even standards for the operation of nuclear reactors), and so on. And in educational settings, it is clear that teachers make arbitrary decisions about what to teach in their courses, how to teach their material, and at what pace they should teach. Surely, if teachers are deemed qualified to make these other important decisions, they are equally qualified to set standards or cut-off scores for the monitoring of student progress in their courses. But what if a cut-off score is set too high (or low) or students are misclassified? Through experience with a curriculum, with high quality criterion-referenced tests, and with careful evaluation work, standards that are not "in line" with others can be identified and revised. And for students who are misclassified there are some redeeming features. Those that perform below the standard will be assigned remedial work and the fact that they performed below the cut-off score suggests that they could not be too far above it (this would be true for most of the students about whom false-negative errors are made) and so the review period will not be a total waste of time.

And for those students who are misclassified because they scored above a cut-off score, they will be tested again. It is possible the next time the error will be caught (particularly if the objectives are sequential). A comment by Ebel (1978) is particularly appropriate at this point:

> Pass-fail decisions on a person's achievement in learning trouble some measurement specialists a great deal. They know about errors of measurement. They know that some who barely pass do so only with the help of errors of measurement. They know that some who fail do so only with the hindrance of errors of measurement. For these, passing or failing does not depend on achievement at all. It depends only on luck. That seems unfair, and indeed it is. But, as any measurement specialist can explain, it is also entirely unavoidable. Make a better test and we reduce the number who will be passed or failed by error. But the number can never be reduced to zero. (p. 549)

The consequences of false-positive and false-negative errors with basic skills assessment or high school certification tests are however considerably more serious and so more attention must be given to the design of these testing programs (for example, content covered by the , the timing of tests, and decisions made with the test results). Considerably more effort must also be given to test development, content validation, and setting of standards.

### 3. How should a method be selected?

There are many factors to consider in selecting a method to determine cut-off scores. For example,

1. How important are the decisions?
2. How much time is available?
3. What resources are available to do the job?
4. How capable are the appropriate individuals of applying a particular method successfully?

The most interesting workwe have seen to date regarding the selection of a method was offered by Jaeger (1976). He considers several methods for determining cut-off scores, several approaches for assigning examinees to mastery states, and various threats to the validity of assignments. While Jaeger's work is theoretic, it provides an excellent starting point for anyone interested in initiating research on the merits of different methods. One thing seems clear from his work—all of the methods he studied appear to have numerous potential drawbacks and so the selection of a method in a given situation should be made carefully.

## 4. What guidelines are available for applying particular methods successfully?

Unfortunately, there are relatively few sets of guidelines available for applying any of the methods. In our judgment, Zieky and Livingston (1977) have provided a very helpful set of guidelines for applying several methods (the popular Nedelsky method and the Angoff method are two of the methods included). Some new work by Popham (1978) is also very helpful. More materials of this type and quality are needed. Some procedural steps for standard-setting with respect to three important uses of tests — (1) daily classroom assessment, (2) basic skills assessment for yearly promotions and high school certification, and (3) professional licensing and certification are provided in section 6.9.

## 6.3 Distinction Between Continuum and State Models

The basic difference between continuum and state models has to do with the underlying assumption made about ability. According to Meskauskas, two characteristics of continuum models are:

1. Mastery is viewed as a continuously distributed ability or set of abilities.
2. An area is identified at the upper end of this continuum, and if an individual equals or exceeds the lower bound of this area, he/she is termed a master.

State models, rather than being based on a continuum of mastery, view mastery as an all-or-none proposition (i.e., either you can do something or you cannot). Three characteristics of state models are:

1. Test true-score performance is viewed as an all-or-nothing state.
2. The standard is set at 100%.
3. After a consideration of measurement errors, standards are often set at values less than 100%.

There are at least three methods for setting standards that are built on a state model conceptualization of mastery. The models take into account measurement error, deficiencies of the examination, etc., in "tempering" the standard from 100%. These methods have been referred to by Glass (1978a) in his review of methods for setting standards as "counting backwards from 100%." State model methods advanced to date include the mastery testing evaluation model of Emrick (1971), the true-score model of Roudabush (1974), and some recently advanced statistical models of Macready and Dayton (1977). However, since state models are somewhat less usefulness than continuum models in elementary and secondary school testing programs, they will not be considered further here. Our failure to consider them further however, should not be interpreted as a criticism of this general approach to standard-setting. The approach seems to be especially applicable with many performance tests (Hambleton & Simon, in preparation).

## 6.4 Traditional and Normative Procedures

Before discussing the various continuum models of standard setting, two other models for standard-setting should be mentioned. These methods, which seem to have limited value in setting standards, have been referred to by a variety of names. We will call them "traditional standards" and "normative standards."

Traditional standards are standards that have gained acceptance because of their frequent use. Classroom examples include the decision that 90 to 100 percent is an A, 80 to 89 percent is a B, etc. It appears that such methods have been used occasionally in setting standards.

"Normative" standards refer to any of three different uses of normative data, two of which are, at best, questionable. In the first method, use is made of the normative performance of some external "criterion" group. As an example, Jaeger (1978) cites the use of the Adult Performance Level (APL) tests by Palm Beach County, Florida schools. Test performance of groups of "successful" adults were used to set standards for high school students. Such a procedure can be criticized on a number of grounds. Jaeger (1978) points out that society changes, and that standards should also change. Standards based on adult performance may not be relevant to high school students. Shepard (1976) points out that any normatively-determined standard will

immediately result in a multitude of counterexamples. Further, Burton (1978) suggests that relationships between skills in school subjects and later success in life are not readily determinable, hence, observing the degree of achievement on the test of some "successful" norm group makes little sense. Jaeger (1978) goes on to say: "There are no empirically tenable survival standards on school-based skills that can be justified through external means."

A second way of proceeding with normative data is to make a decision about a standard based solely on the distribution of scores of examinees who take the test. Such a procedure circumvents the "minimum test score for success in life" problem, but the procedure is still not useful for setting standards. For example, Glass (1978a) cites the California High School Proficiency Examination, where the 50th percentile of graduating seniors served as the standard. What can be said of a procedure where whether or not an individual passes or fails a minimum competency test depends upon the other individuals taking the test? In the California situation, the standard was set with no reference at all to the content of the test or the difficulty of the test items.

The third use of normative data discussed in the literature concerns the supplemental use of normative data in setting a standard. Shepard (1976), Jaeger (1978), and Conaway (1976, 1977) all favor such a procedure. Recently Jaeger (1978) advanced a standard setting method which requires judges to make judgments partially on the basis of item content. In his method, Jaeger calls for incorporation of some tryout test data

to aid judges in reconsidering their initial assessments. Shepard (1976) makes the following point:

> Expert judges ought to be provided with normative data in their deliberations. Instead of relying on their experience, which may have been with unusual students or professionals, experts ought to have access to representative norms. . .of course, the norms are not automically the standards. Experts still have to decide what "ought" to be, but they can establish more reasonable expectations if they know what current performance is than if they deliberate in a vacuum.

We agree with Jaeger, Conaway, and Shepard about the usefulness of normative data when used in conjunction with a standard setting method.

## 6.5 Consideration of Several Promising Standard Setting Methods

Remaining methods for setting standards to be discussed in this unit assume that domain score estimates derived from criterion-referenced tests are on a continuous scale (hence, the methods fall under the heading of "Continuum Model"). For convenience, the methods under discussion are organized into three categories. The methods are presented in Figure 6.5.1. The categories are labelled "judgmental," "empirical," and "combination." In judgmental methods, data are collected from judges for setting standards, or judgments are made about the presence of variables (for example, guessing) that would effect the placement of a standard. Empirical methods require the collection of examinee response data to aid in the standard-setting process. Combination methods, not surprising, incorporate judgmental data and empirical data into the standard-setting process.

Figure 6.5.1 A classification of methods for setting standards[2]

| Judgmental Methods | | Combination Methods | | Empirical Methods[1] | |
|---|---|---|---|---|---|
| Item Content | Guessing | Judgmental-Empirical | Educational Consequences | Data—Two Groups | Data-Criterion Measure |
| Nedelsky (1954) | Millman (1973) | Contrasting Groups (Zieky and Livingston, 1977) | Block (1972) | Berk (1976) | Livingston (1975) |
| Modified Nedelsky (Nassif, 1978) | | Borderline Groups (Zieky and Livingston, 1977) | | | Livingston (1976) |
| Angoff (1971) | | | | | Huynh (1976) |
| Modified Angoff (ETS, 1976) | | | | | Van der Linden and Mellenbergh (1977) |
| Ebel (1972) | | | | | |
| Jaeger (1978) | | | | | |
| | | Bayesian Methods | | Decision-Theoretic | |
| | | Hambleton and Novick (1973) | | Kriewall (1972) | |
| | | Novick, Lewis, Jackson (1973) | | | |
| | | Schoon, Gullion Ferrara (1978) | | | |

[1]Involve the use of examinee response data.

[2]From a paper by Hambleton and Eignor (1979).

## 6.6 Judgmental Methods

### 6.6.1 Item Content

In this situation, individual items are inspected, with the level of concern being how the minimally competent person would perform on the items. In other words, a judge is asked to assess how or to what degree an individual who could be described as minimally competent would perform on each item. It should be noted before describing particular procedures utilizing this criterion that while this is a good deal more objective than setting standards based on any of the methods previously discussed, a considerable degree of subjectivity still exists. Six procedures based on item content assessment will now be discussed.

#### i. Nedelsky Method

In Nedelsky's method, judges are asked to view each question in a test with a particular criterion in mind. The criterion for each question is, which of the response options should the minimally competent student (Nedelsky calls them "D-F students") be able to eliminate as incorrect? The minimum passing level (MPL) for that question then becomes the reciprocal of the remaining alternatives. For instance, if on a five-alternative multiple choice question, a judge feels that a minimally competent person could eliminate two of the options, then for that question, MPL = $\frac{1}{3}$. The judges proceed with each question in a like fashion, and upon completion of the judging process, sum the values for each question to obtain a standard on the total set of test items. Next, the individual judge's standards are averaged. The average is denoted $\hat{\pi}_o$.

Nedelsky felt that if one were to compute the standard deviation of individual judge's standards, this distribution would be synonomous with

the (hypothesized or theoretical) distribution of the scores of the borderline students. This standard deviation, $\sigma$, could then be multiplied by a constant K, decided upon by the test users, to regulate how many (as a percent) of the borderline students pass or fail. The final formula then becomes:

$$\hat{\pi}_o = \hat{\pi}_o + K\,\sigma$$

How does the $K\,\sigma$ term work? Assuming an underlying normal distribution, if one sets K=1, then 84% of the borderline examinees will fail. If K=2, then 98% of these examinees will fail. If K=0, then 50% of the examinees on the borderline should fail. The value for K is set by (say) a committee prior to the examination.

The final result of the application of Nedelsky's method will be an absolute standard. This is because the standard is arrived at without consideration of the score distributions of any reference group. In fact, the standard is arrived at prior to using the test with the group one is concerned with testing.

The following example is included to demonstrate how the Nedelsky method can be applied in a criterion-referenced testing situation.

Example: Suppose five judges were asked to score, using the Nedelsky method, a six question criterion-referenced test made up of questions that have five response options each. Further, suppose the judges agreed that they would like 84% of the "D-F" or minimally competent students to fail (i.e., they set K=+1). The calculations below show the steps necessary to calculate a cut-off score for the test.

| Judge | Test Item 1 | 2 | 3 | 4 | 5 | 6 | Cut-Off Score from Each Judge |
|---|---|---|---|---|---|---|---|
| A | .25 | .33 | .25 | .25 | .00 | .33 | 1.41 |
| B | .25 | .50 | .25 | .50 | .25 | .33 | 2.08 |
| C | .33 | .33 | .25 | .33 | .25 | .33 | 1.82 |
| D | .25 | .33 | .25 | .33 | .25 | .33 | 1.74 |
| E | .00 | .50 | .25 | .33 | .00 | .25 | 1.33 |

$$\text{Average Cut-Off Score (Across Five Judges)} = \frac{1.41 + 2.08 + 1.81 + 1.74 + 1.33}{5}$$

$$= 1.68$$

$$\text{Standard Deviation of the Cut-Off Scores} = \sqrt{\frac{(1.41-1.68)^2+(2.08-1.68)^2+\ldots+(1.33-1.68)^2}{5}}$$

$$= \sqrt{\frac{.380}{5}}$$

$$= .28$$

$$\text{Adjusted Cut-Off Score (84\% of Borderline Student to Fail)} = 1.68 + 1 \times .28$$

$$= 1.96$$

Therefore, approximately two test items out of six is the cut-off score on this test. From a practical standpoint, this value would seem low, but the data is created to demonstrate the process and not to model a real testing situation. Therefore, no practical significance should be attached to the answer.

ii. Modified Nedelsky

Nassif (1978), in setting standards for the competency-based teachers education and licensing systems in Georgia, utilized a modified Nedelsky procedure. A modification of the Nedelsky method was needed to handle the volume of items in the program. In the modified Nedelsky task, the entire item (rather than each distractor) is examined and classified in terms of two levels of examinee competence. The following question was asked about each item: "Should a person with minimum competence in the teaching field be able to answer this item correctly?" Possible answers were "yes," "no," and "I don't know." Agreement among judges can be studied through a simple comparison of the ratings judges give to each item. A standard may be obtained by computing the average number of "yes" responses judges give to the entire set of test items.

iii. Ebel's Method

Ebel (1972) goes about arriving at a standard in a somewhat different manner, but his procedure is also based upon the test questions rather than an "outside" distribution of scores. Judges are asked to rate items along two dimensions: Relevance and difficulty. Ebel uses four categories of relevance: Essential, important, acceptable and questionable. He uses three difficulty levels: Easy, medium and hard. These categories then form (in this case) a 3 x 4 grid. The judges are next asked to do two things:

1. Locate each of the test questions in the proper cell, based upon relevance and difficulty,
2. Assign a percentage to each cell; that percentage being the percentage of items in the cell that the minimally-qualified examinee should be able to answer.

Then the number of questions in each cell is multiplied by the appropriate percentage (agreed upon by the judges), and the sum of all the cells, when divided by the total number of questions, yields the standard.

The example that follows is modeled after an example offered by Ebel (1972).

Example: Suppose that for a 100 item test, five judges came to the following agreement on percentage of success for the minimally qualified candidate.

| Relevance | Difficulty Level | | |
|---|---|---|---|
| | Easy | Medium | Hard |
| Essential | 100%* | 80% | -- |
| Important | 90% | 70% | -- |
| Acceptable | 90% | 40% | 30% |
| Questionable | 70% | 50% | 20% |

*The expected percentage of passing for items in the category.

Combining this data with the judges location of test questions in the particular cells would yield a table like the following:

| Item Category | Number of Items* | Expected Success | Number X Success |
|---|---|---|---|
| ESSENTIAL | | | |
| Easy | 85 | 100 | 8500 |
| Medium | 55 | 80 | 4400 |
| IMPORTANT | | | |
| Easy | 123 | 90 | 11070 |
| Medium | 103 | 70 | 7210 |
| ACCEPTABLE | | | |
| Easy | 21 | 90 | 1890 |
| Medium | 43 | 40 | 1720 |
| Hard | 50 | 30 | 1500 |
| QUESTIONABLE | | | |
| Easy | 2 | 70 | 140 |
| Medium | 8 | 50 | 400 |
| Hard | 10 | 20 | 200 |
| TOTAL | 500 | | 37030 |

$$\frac{37030}{500} = 74$$

*The number of items placed in each category by all five of the judges.

Three comments can be made about Ebel's method that should be sufficient to suggest caution when using it. One, Ebel offers no prescription for the number or type of descriptions to be used along the two dimensions. This is left to the judgment of the individuals judging the items. It is likely that a different set of descriptions applied to the same test would yield a different standard. Two, the process is based upon the decisions of judges, and while the standard could be called absolute, in that it is not referenced to score distribution, it can't be called an "objec-

tive" standard. Three, a point about Ebel's method has been offered by Meskauskas (1976):

> In Ebel's method, the judge must simulate the decision process of the examinee to obtain an accurate judgment and thus set an appropriate standard. Since the judge is more knowledgeable than the minimally-qualified individual, and since he is not forced to make a decision about each of the alternatives, it seems likely that the judge would tend to systematically over-simplify the examinees task . . . Even if this occurs only occasionally, it appears likely that, in contrast to the Nedelsky method, the Ebel method would allow the raters to ignore some of the finer discriminations that an examinee needs to make and would result in a standard that is more difficult to reach. (p. 138)

iv. Angoff's Method

When using Angoff's technique, judges are asked to assign a probability to each test item directly, thus circumventing the analysis of a grid or the analysis of response alternatives. Angoff (1971) states:

> . . .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 proportions, would then represent the minimally acceptable score. (p. 515)

v. Modified Angoff

ETS (1976) utilized a modification of Angoff's method for setting standards. Based on the rationale that the task of assigning probabilities may be overly difficult for the items to be assessed (National Teacher Exams) Educational Testing Service instead supplied a seven point scale on which certain percentages were

fixed. Judges were asked to estimate the percentage of minimally knowledgeable examinees who would know the answer to each test item. The following scale was offered:

5 20 40 60 75 90 95 DNK

where "DNK" stands for "Do Not Know."

ETS has also used scales with the fixed points at somewhat different values; the scales are consistent though in that seven choice points are given. For the Insurance Licensing Exams, 60 was used as the center point, since the average percent correct on past exams centered around 60%. The other options were then spaced on either side of 60.

vi. Jaeger's Method

Jaeger (1978) recently presented a method for standard-setting on the North Carolina High School Competency Test. Jaeger's method incorporates a number of suggestions made by participants in a 1976 NCME annual meeting symposium presented in San Francisco by Stoker, Jaeger, Shepard, Conaway, and Haladyna; it is iterative, uses judges from a variety of backgrounds, and employs normative data. Further, rather than asking a

question involving "minimal competence," a term which is hard to operationalize, and conceptualize, Jaeger's questions are instead:

> "Should every high school graduate be able to answer this item correctly?" "____Yes, ____No." and
> "If a student does not answer this item correctly, should he/she be denied a high school diploma?"
> "____Yes, ____No."

After a series of iterative processes involving judges from various areas of expertise, and after the presentation of some normative data, standards determined by all groups of judges of the same type are pooled, and a median is computed for each type of judge. The minimum median across all groups is selected as the standard.

## Comparisons Among Judgmental Models

We are aware of two studies that compare judgmental methods of setting standards; one study was done in 1976, the other is presently underway at ETS.

In 1976, Andrew and Hecht carried out an empirical comparison of the Nedelsky and Ebel methods. In that study, judges met on two separate occasions to set standards for a 180 item, four options per item, exam to certify professional workers. On one occasion the Nedelsky method was used. On a second occasion the Ebel method was used. The percentage of test items that should be answered correctly by a minimally competent examinee was set at 69% by the Ebel method and at 46% by the Nedelsky method.

Glass (1978a) described the observed difference as a "startling finding". Our view is that since directions to the judges were different, and procedures differed, we would not expect the results from these two methods to be similar. The authors themselves report:

> It is perhaps not surprising that two procedures which involve different approaches to the evaluation of test items would result in different examination standards. Such examination standards will always be subjective to some extent and will involve different philosophical assumptions and varying conceptualizations. (p. 49)

Ebel (1972) makes a similar point:

> . . .it is clear that a variety of approaches can be used to solve the problem of defining the passing score. Unfortunately, different approaches are likely to give different results. (p. 496)

Possibly the most important result of the Andrew-Hecht study was the high level of agreement in the determination of a standard using the same method across two teams of judges. The difference was not more than 3.4% within each method. Data of this kind address a concern raised by Glass (1978a) about whether judges can make determinations of standards consistently and reliably. At least in this one study, it appears that they could. From our interactions with staff at ETS who conduct teacher workshops on setting standards, we have learned that teams of teachers working with a common method obtain results that are quite similar. And this result holds across tests in different subject matter areas and at different grade levels. We have observed the same result in our own work. Of course, certain conditions must be established if agreement among judges is to be obtained. Essentially, it is necessary that the judges share a common definition of the "minimally competent" student and fully understand the rating process they are to use.

6.6.2 Guessing and Item Sampling

In this section, some concerns initially expressed by Millman (1973) about errors due to guessing and item sampling will be discussed.

If the test items allow a student to answer questions correctly by guessing, a systematic error is introduced into student domain score estimates. There are three possible ways to rectify this situation:

1. The cut-off score can be raised to take into account the contribution expected from the guessing process.
2. A student's score can be corrected for guessing and then the adjusted score compared to the performance standard.
3. The test itself can be constructed to minimize the guessing process.

Methods one and two assume that guessing is of a pure, random nature, which is not likely to be the case for criterion-referenced tests. Thus, adjusting either the cutting score or the student's scores will probably prove to be inadequate. The test must be structured to keep guessing to a minimum, because if it occurs, it can't be adequately corrected for.

Also, if because of problems of test construction, inconvenience of administration, or a host of other problems, the test is not representative of the content of the domain, then Millman (1973) suggests that the cutting score or standard be raised (or lowered) an amount to protect against misclassification of students; i.e., false-positive and false-negative errors. Millman offers no methods for determining the extent or direction of correction for these problems. We feel that the test practitioner should exert extra effort to assure that the problem just discussed doesn't occur in the first place. Once again, there doesn't appear to be an adequate method for "correcting away" the problem.

## 6.7 Empirical Methods

### 6.7.1 Data From Two Groups

Berk (1976) presented a method for setting cut-off scores that is based on empirical data. He selects empirically the optimal cutting score for a test based upon test data from two samples of students, one of which has been instructed on the material, and the other uninstructed. Before discussing his methodology, where he offers three ways of proceeding based upon the data collected, it is worth discussing why he chose to formulate his model in the first place. He suggests that the extant approaches of a nature similar to his, namely those based on the binomial distribution and those based upon Bayesian decision-theory, suffer from a deficiency. According to Berk:

> The fundamental deficiency of all of these methods is their failure to define mastery operationally in terms of observed student performance, the objective or trait being measured, and item and test characteristics. The criterion level or cutting score is generally set subjectively on the basis of "judgment" or "experience" and the probabilities of Type I/Type II classification errors associated with the criterion are estimated.

One of Berk's procedures considers false-positive and false-negative errors, but the difference is that the results are based upon actual data.

Berk offers three ways of approaching the problem of setting standards utilizing empirical data: (1) Classification of outcome probabilities, (2) computation of a validity coefficient, and (3) utility analysis.

#### i. The Basic Situation

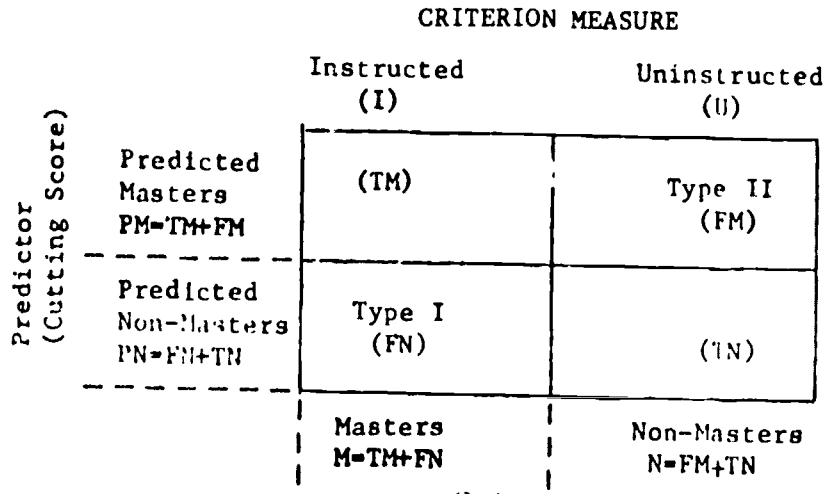
Two criterion groups are selected for use in this procedure, one group comprised of instructed students and another of uninstructed students. The instructed group should, according to Berk, "consist of those students who have received 'effective' instruction on the objective to be assessed."

Berk suggests that these groups should be approximately equal in size and large enough to produce stable estimates of probabilities. Test items measuring one objective are then administered to both groups and the distribution of scores (putting both groups together) can be divided by a cut-off score into two categories.

Combining the classifications of students by predictor (test score) and criterion (instructed vs. non-instructed status) results in four categories that can be represented in a 2 x 2 table, with relevant marginals:

1. True Master (TM): an instructed student whose test score is above the cutting score (C).
2. False Master (FM): A Type II misclassification error where an uninstructed student's test score lies above the cutting score (C).
3. True Non-Masters (TN): An uninstructed student whose test score lies below the cutting point (C).
4. False Non-Masters (FN): Type I misclassification where an instructed student's test score lies below C.

Tabularly, this can be presented as follows. Note how the marginals are defined because they are used in the formulations to follow.



## ii. Classification of Outcome Probabilities

In this procedure, identification of the optimal cutting score involves an analysis of the two-way classification of outcome probabilities shown above. This can be done algebraically by following the steps listed below, or graphically, as illustrated in a subsequent section. The steps to follow are:

1. Set up a two-way classification of the frequency distribution for each possible cutting score.
2. Compute the probabilities of the 4 outcomes (for each cutting score) by expressing the cell frequencies as proportions of the total sample. For instance

   Prob (TM) = TM/(M+N)

   Prob (FM) = FM/(M+N)

   Prob (TN) = TN/(M+N)

   Prob (FN) = FN/(M+N)

3. For each cutting score, add the probability of correct decisions: Prob (TM) + Prob (TN), and the probability of incorrect decisions: Prob (FN) + Prob (FM).
4. The optimal cutting score is the score that maximizes Prob (TM) + Prob (TN) and minimizes Prob (FN) + Prob (FM). It is sufficient to observe the score that maximizes Prob (TM) + Prob (TN) because [Prob (FN) + Prob (FM)] = 1 - [Prob (TM) + Prob (TN)]. That is, the score that maximizes the probability of correct decisions automatically minimizes probability of incorrect decisions.

iii. Graphical Solution

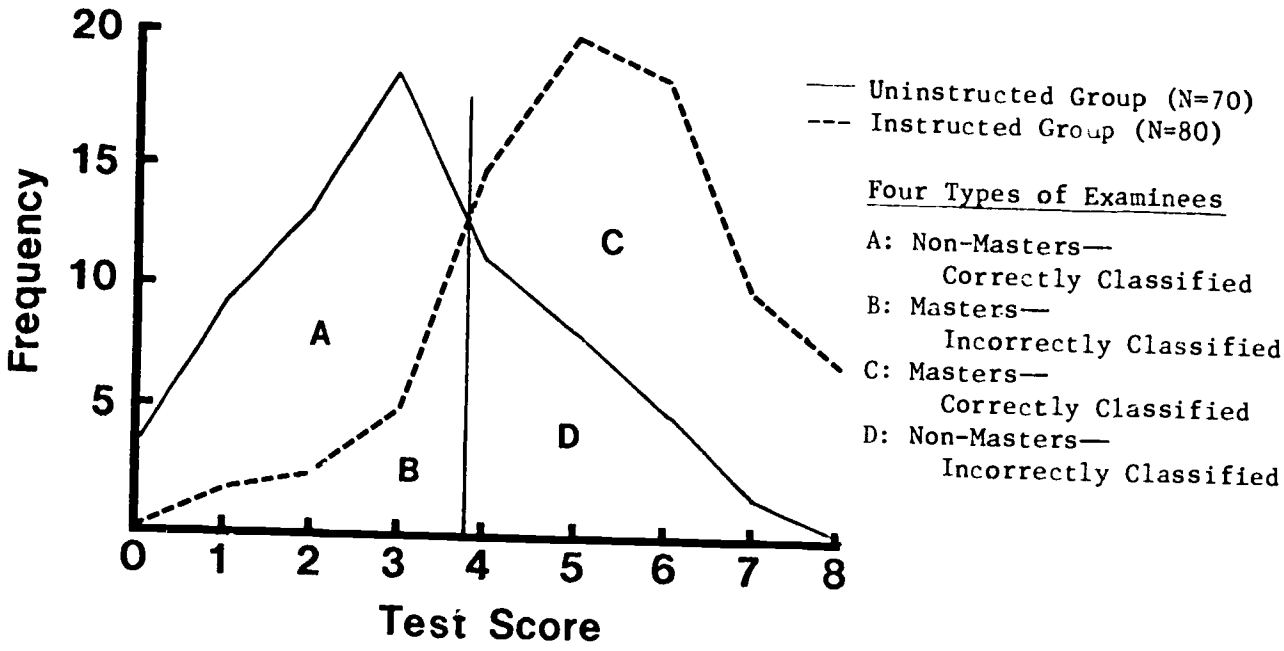
Berk (1976) also mentions that the optimal cutting point for a criterion-referenced test can be located by observing the frequency distributions for the instructed and uninstructed groups. According to Berk:

> The instructed and uninstructed group score distributions are the primary determinants of the extent to which a test can accurately classify students as true masters and true non-masters of an objective. The degree of accuracy is, for the most part, a function of the amount of overlap between the distribution.

If the test distributions overlap, no decisions can be made. The ideal situation would be one in which the two distributions have no overlap at all. A typical situation we should hope for is for the instructed group distribution to have a negative skew, the uninstructed group to have a positive skew, and for there to be a minimum of overlap. The point at which the distributions intersect is then the optimal cut-off score.

In Figure 6.7.1, the distributions of test scores for two groups of examinees (one instructed group and one uninstructed group) are shown.

Figure 6.7.1 Frequency polygons of criterion-referenced test scores for two groups - an instructed group and an uninstructed group on the content measured by the test.




**Frequency Distribution of Test Scores**

| Test Score | U I Group | I Group |
|---|---|---|
| 8 | 0 | 7 |
| 7 | 2 | 10 |
| 6 | 5 | 18 |
| 5 | 8 | 20 |
| 4 | 11 | 15 |
| 3 | 18 | 5 |
| 2 | 13 | 3 |
| 1 | 9 | 2 |
| 0 | 4 | 0 |

iv. Validity Coefficient

In this procedure, a validity coefficient is computed for each possible cutting score. The cutting score yielding the highest validity coefficient also yields the highest probability of correct decisions. To utilize the procedure, the following steps should be followed:

1. From the two-way classification introduced earlier, compute the base rate (BR) and the selection ratio (SR). They are given by:

   BR = Prob (FN) + Prob (TM)

   SR = Prob (TM) + Prob (FM)

2. Calculate the phi coefficient $\phi_{vc}$ using the following formula:

$$\phi_{vc} = \frac{\text{Prob (TM)} - \text{BR (SR)}}{\sqrt{\text{BR (1-BR) SR(1-SR)}}}$$

3. The cutting score yielding the highest $\phi_{vc}$ is the optimal cutting score.

The formula for the phi coefficient, $\phi_{vc}$, given above is suitable for a 2 x 2 table of cell probabilities. More generally, the phi coefficient is the Pearson product moment correlation between two dichotomous variables, and could be arrived at as follows:

1. Each student with a test score above the cutting score in question is assigned a 1, below a 0.
2. Each student in the instructed group is assigned a 1, in the uninstructed group, a 0.
3. $\phi_{vc}$ would then be the correlation coefficient computed in the usual way.

v. Utility Analysis

In this section, costs or losses are assigned to the misclassification of students as false masters or false non-masters. The procedures here are closely tied to the decision-theoretic procedures discussed in a later section. The procedure is presented at this point because it can be related to the two Berk procedures just discussed.

First of all, Berk notes the following fact.

> When the outcome probabilities or validity coefficient approach is used to select the optimal cutting score, it is assumed that the 2 types of errors are equally serious. If, however, this assumption is not realistic in terms of the losses which may result from a particular decision, the error probabilities need to be weighted to reflect the magnitude of he losses associated with the decision.

Berk notes that determination of the relative size of each loss is judgmental, and must be guided by the consequences of the decision considered. He mentions considering the following factors: Student motivation, teacher time, availability of instructional materials, content, and others. Berk suggests the following, which we have capsulized into a series of steps:

1. Estimate the expected disutility of a decision strategy ($\zeta$) by

   $\zeta_k$ = Prob (FN)[$D_1$] + Prob (FM) [$D_2$]

   where $D_1$ and $D_2 < 0$

   and k = the single decision in question

   $D_1$ and $D_2$ = respective disutility values

2. Estimate the expected utility of a decision strategy ($v$) by

   $v_k$ = Prob (TM) [$U_1$] + Prob (TN) [$U_2$]

   where $U_1$ and $U_2 > 0$

   and k = the single decision in question (same as for disutility)

   $U_1$ and $U_2$ = respective utility values

3. Form a composite measure of test usefulness by combining the estimates of utility and disutility across all decisions

$$\gamma = \sum_{k=1}^{n} (\nu_k + \zeta_k)$$

$\gamma$ = index of expected maximal utility.

4. Choose the cutting score with the highest $\gamma$ index (it maximizes the usefulness of the test for decisions with a specific set of utilities and disutilities).

vi. Suggestions

The procedures developed by Berk (1976) hold considerable promise for use in setting criterion-referenced test score standards. The ideas in his procedures are now new; there are other procedures that are concerned with the maximization of correct decisions and the minimization of false-positive and false-negative errors. The attractive feature is the ease with which Berk's methods can be understood and applied. The major potential drawback is in the assignment of examinees to criterion groups. If many examinees in the "instructed group" do not possess the assumed knowledge and skills measured by the criterion-referenced test (or if many examinees in the "uninstructed group" do), Berk's methods will produce inaccurate results.

6.7.2 Decision-Theoretic Procedures

Berk (1976) looked at the minimization of false-positive and false-negative decisions through the use of actual test data. He selects as optimal the cutting score that minimizes false-positive and false-negative errors. Another way to look at false-positive and false-negative errors is to assume an underlying distributional form for your data and then

observe the consequences of setting values, such as cutting points, based upon the distributional model. The logic is the same here in terms of minimization of errors, except that by assuming a distributional form, actual data does not have to be collected. Situations can be simulated or developed, based upon the model.

Meskauskas (1976) has related and compared these procedures to those based upon analyses of the content of the test. In reference to these models, of which we will describe one:

> . . .the models to follow deal with approaches that start by assuming a standard of performance and then evaluating the classification errors resulting from its use. If the error rate is inappropriate, the decision maker adjusts the standard a bit and tries his equation again.

Before discussing one of the procedures in greater detail, the Kriewall binomial-based model, the procedures discussed here should be related to criterion-referenced testing procedures involving the determination of test length. Many of the test length determination procedures (Millman, 1973; Novick & Lewis, 1974) make underlying distributional assumptions and proceed in the fashion discussed above by Meskauskas. The focus of concern, however, is test length determination, and not the setting of a cutting score. In fact, Millman's (1973) procedure is based upon exactly the same underlying distribution, the binomial, as is Kriewall's model to be discussed. It should be pointed out that the procedures are exactly the same, the data is just represented differently because of the level of concern, either cutting score or test length.

i. Kriewali's Model

Kriewall's (1972) model focuses on categorization of learners into several categories: Non-master, master, and an in-between state where the student has developed some skills, but not enough to be considered a master.

Kriewall assumes the func ion of measurement, using the test, is to classify students into one of two categories, master or non-master. Of course, the test, as a sample of the domain of tasks, is going to misclassify some individuals as false-positives (masters based on the test, but non-masters in reality) and false-negatives (non-masters on the test, but masters in reality). By assuming a particular distribution, these errors may be studied.

Kriewall's probability model, used to develop the likelihood of classification errors, is based upon the binomial distribution. He assumes:

1. The test represents a randomly selected set of dichotomously scored (0-1) items from the domain.
2. The likelihood of correct response for a given individual is a fixed quantity for all items measuring a given objective.
3. Responses to questions by an individual are independent. That is, the outcome of one trial (taking one question) is independent of the outcome of any other trial.
4. Any distribution of difficulty of questions (for an individual) within a test is assumed to be a function of randomly occurring erroneous responses (Meskauskas, 1976).

With these assumptions, Kriewall views a student's test performance as "a sequence of independent Bernoulli trials, each having the same probability of success." A sequence of Bernoulli trials follows a binomial distribution, which has a probability function which relates the probability of occurrence of an event (a particular test score) to the number of questions in the test by:

$$f(x) = \binom{n}{x} p^x q^{n-x} \quad ,$$

where

x = a test score

n = total number of test items

p = examinee domain score

q = 1-p

and

$$\binom{n}{x} \quad \frac{n!}{x!\,(n-x)!} \quad .$$

Kriewall sets some boundary values and a cutting score, and then looks at the probability of misclassification errors. Using the notation of Meskauskas (1976), set:

$z_1$ = the lower bound of the mastery range (as a proportion of errors)

$z_2$ = the upper bound of the non-mastery range

C = the cutting score; the maximal number of allowable errors for masters. Kriewall recommends $C = \frac{z_1 + z_2}{2}$

Given values for the above three variables, Kriewall uses the (assumed) binomial distribution to determine the probabilities. If $\alpha$ is the probability of a false positive result (a non-master who scores in the mastery category)

and β is the probability of a false negative result (a master who scores in the non-mastery category), then α and β are given by:

$$\alpha = \sum_{w=c}^{n} \binom{n}{w} z_1^{n-w} (1 - z_1)^w$$

$$\beta = \sum_{w=o}^{c-1} \binom{n}{w} z_2^{n-w} (1 - z_2)^w$$

where w = observed number of errors (and w = n-x) for an individual.

According to Meskauskas (1976) the formula for α is:

> . . .equivalent to obtaining the probability that, given a large number of equivalent trials, a person whose true score is equal to the lowest score in the mastery range will fall in the non-mastery range.

By setting $z_1$ and $z_2$ at various values, and determining $C = \frac{z_1 + z_2}{2}$, the probabilities of false positive and false negative errors can be studied. The optimal value for C (and thus $z_1$ and $z_2$) would then be the value that minimized α and β. The results are dependent, however, on n and w.

ii. Suggestions

While Kriewall has offered a method of studying classification errors that does not depend upon actual data, we prefer the method of Berk, due to its simplicity. Kriewall's model seems to us to fit in much better with the procedures on test length determination. For instance, suppose you have specified minimal values for α and β, and have determined C, the cutting point. Then the formulas above for α and β can be solved for n, the total number of questions needed. (It would be much easier if one isolated n on the left hand side). This is exactly what is done when using the binomial model to solve the test length problem.

In sum, we prefer the Berk method for observing probabilities of misclassification errors both because of its simplicity and because of the lack of restricting underlying distributional assumptions. Kriewall's method does, however, offer a viable alternative for setting a cut-off score when actual test data cannot be collected.

### 6.7.3 Empirical Models Depending Upon a Criterion Measure

The models to be discussed in this section bear great resemblance to both Berk's and Kriewall's methods just discussed. They have been separated from those two methods because these methods are built upon the existence of an outside criterion measure, performance measure, or true ability distribution. The test itself, and the possible cut-off scores, are observed in relationship with this outside measure. The optimal cut-off is then chosen in reference to the criterion measure. For instance, Livingston's (1975) utility-based approach leads to the selection of a cut-off score that optimizes a particular utility function. The procedure of Vander Linden and Mellenburgh (1976), in contrast, leads to the selection of a cut-off score that minimizes expected loss.

In reference to the setting of performance standards based upon benefit (and cost) Millman (1973) has suggested that psychological and financial costs be considered:

> All things being equal, a low passing score should be used when the psychological and financial costs associated with a remedial instructional program are relatively high. That is, there should be fewer failings when the costs of failing are high. These "costs" might include lower motivation and boredom,

> damage to self-concept, and dollar and time expenses of conducting a remedial instructional program. A higher passing score can be tolerated when the above costs are not too great or when the negative effects of moving a student too rapidly through a curriculum (i.e., confusion, inefficient learning and so forth) are seen as very important to avoid.

In sum, to utilize these procedures, a suitable outside criterion measure must exist. Success and failure (or probability of success and failure) is then defined on the criterion variable and the cut-off chosen as the score on the test that maximizes (or minimizes) some function of the criterion variable. The existence of such a criterion variable has implications for the utilization of these methods for setting cut-off scores on minimum competency tests.

### i. Livingston's Utility-based Approach

Livingston (1975) suggests the use of a set of linear or semi-linear utility functions in viewing the effects of decision-making accuracy based upon a particular performance standard or cut-off score. That is, the functions relating benefit (and cost) of a decision are related linearly to the cutting score in question.

Livingston's procedure is like Berk's procedure for utility analysis discussed in 6.7.1 except that Livingston develops his procedure based upon any suitable criterion measure (not just instructed versus uninstructed), and also specifies the relationship between utility (benefit or loss) and cutting scores as linear. The relationship does not have to be linear; however, using such a relationship simplifies matters somewhat. In such a situation the cost (of a bad decision) is proportional to the size of the errors made and the benefit (of a good decision) is proportional to the size of the errors avoided.

ii. Van der Linden and Mellenburgh's Approach

The developers of this procedure have prescribed a method for setting cutting scores that is related both to Berk's procedure and Livingston's. We will describe the procedure briefly and in the process relate it to Berk's work. A test score is used to classify examinees into two categories: Accepted (scores above the cutting score) and rejected (scores below). Also, a latent ability variable is specified in advance and used to dichotomize the student population: Students above a particular point on the latent variable are considered "suitable" and below "not suitable." The situation may be represented as follows.

| | | Latent Variable: Not suitable $\gamma < d$ | Latent Variable: Suitable $\gamma \geq d$ |
|---|---|---|---|
| Decision | Accepted $X \geq C$ | "False +" $\ell_{01}(\gamma)$ | $\ell_{11}(\gamma)$ |
| | Rejected $X < C$ | $\ell_{00}(\gamma)$ | "False -" $\ell_{10}(\gamma)$ |

where C = cutting score on the criterion-referenced test

d = cutting score on the latent variable ($0 \leq d \leq 1$),

and where $\ell_{ij}$ $(i,j = 0,1)$ is a function of $\gamma$ and related in the general loss function:

$$L = \begin{cases} \ell_{00}(\gamma) \text{ for } \gamma < d,\ X < C \\ \ell_{10}(\gamma) \text{ for } \gamma \geq d,\ X < C \\ \ell_{01}(\gamma) \text{ for } \gamma < d,\ X \geq C \\ \ell_{11}(\gamma) \text{ for } \gamma > d,\ X \geq C \end{cases}$$

The authors then specify risk (the quantity to be minimized) as the expected loss, and the cutting score that is optimal is the value of C that minimizes the risk function (expected value of loss). They simplify matters (as does Livingston) by specifying their loss function as linear.

In sum, while Van der Linden and Mellenburgh have provided a method for setting a cut-off score on the test, they have offered little to help in setting the cut-off on the latent variable. In a sense then, they have only transferred the problem of setting a standard to a different measure!

### iii. Livingston's Use of Stochastic Approximation Techniques

Livingston (1976) has developed procedures for setting cut-off scores based upon stochastic approximation procedures. According to Livingston, the problem involving cut-off scores can be phrased as follows to fit stochastic procedures: "In general, the problem is to determine what level of input (written test score) is necessary to produce a given response (performance), when measurements of the response are difficult or expensive." The procedure, according to Livingston, is as follows:

1. Select a person; record his/her test score and measure his/her performance.
2. If the person succeeds on the performance measure (if his/her performance is above the minimum acceptable), choose next a person with a somewhat lower test score. If the person fails on the performance measure, choose a person with a higher written test score.
3. Repeat step 2, choosing the third person on the basis of the second person's measured performance.

Livingston offers two different procedures for choosing step size, the up-and-down and the Robbins-Monro Procedure, and a number of procedures for estimating minimum passing scoresconsonant with each.

This procedure, like those discussed earlier in this section, depends upon the existence of a cut-score established on another variable, this time the performance measure, in order to establish the passing score on the test. This then limits greatly the applicability of the method. Livingston (personal communication, 1978) has suggested that judgmental data on performance can be used, rather than actual performance data, with the procedure, but this has yet to be documented in any fashion. When documented, the possibilities for use of the procedures will be greatly expanded.

iv. Huynh's Procedures

Huynh (1976) has advanced procedures for setting cut-off scores that are predicated on the existence of a "referral task." This referral task can be envisioned as an external criterion to which competency can be related. For instance, Huynh (1976) states that "Mastery in one unit of instruction may not be reasonably declared if it cannot be assumed that the masters would have better chances of success in the next unit of instruction." The next unit in this case would be the referral task.

These procedures once again depend upon an outside criterion variable to permit the estimation of a cut-score. In

this case, the user of the method is asked to establish the probability of success of individuals on the referral task. Because of the necessity of a criterion variable for operation, these procedures suffer in generalizability. They are, for instance, apparently not useful for minimum competency testing situations where a criterion variable, and associated probability of success, are next to impossible to establish.

### 6.7.4 Educational Consequences

In this situation, one is concerned with looking at the effect setting a standard of proficiency has on future learning or other related cognitive or affective success criteria. According to Millman (1973), the question here is "What passing score maximizes educational benefits?".

This approach can be visualized from an experimental design point of view. A subject matter domain is taught to a class of students who are then tested on the material. These students are assigned (randomly) to groups with the groups differing on the performance level required for passing the test. The students are then assessed on some valued outcome measure and the level of performance on the criterion-referenced test for which the valued outcome is maximal (it could be a combination of valued outcomes) becomes the performance standard or criterion score.

Thus, to use this method, much more data needs to be collected than for the item content procedures. An experiment must be conducted, and then a cut-off score is selected based upon the results of the experiment.

Because of the difficulties involved in designing and carrying out experiments in school settings, the method is unlikely to find much use.

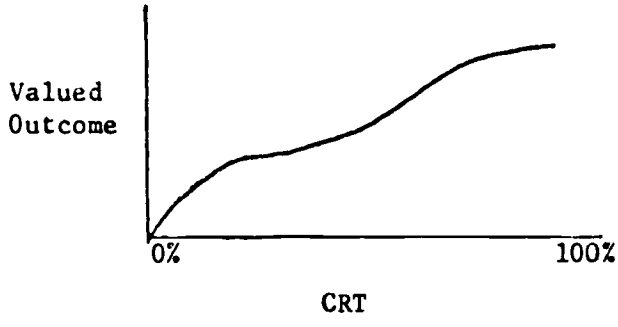
### 1. Block's Study

Block's study (1972) involves students learning a subject segment on matrix algebra using a Mastery Learning paradigm. Such a paradigm dictates that students who don't perform adequately on the posttest be recycled through remedial activities until they demonstrate mastery (re: attain a score above the cutting score). Block established four groups of students, where each group was tested using one of the following four performance standards: 65, 75, 85, and 95% of the material in a unit must be mastered before proceeding on the next unit. He then examined the effects of varying the performance standard on six criteria that were used as the variables to be maximized. Viewing these criteria as either cognitive or affective, Block observed that the 95% performance level maximized student performance on the cognitive criteria, while the 85% performance level seemed to maximize the affective criteria.

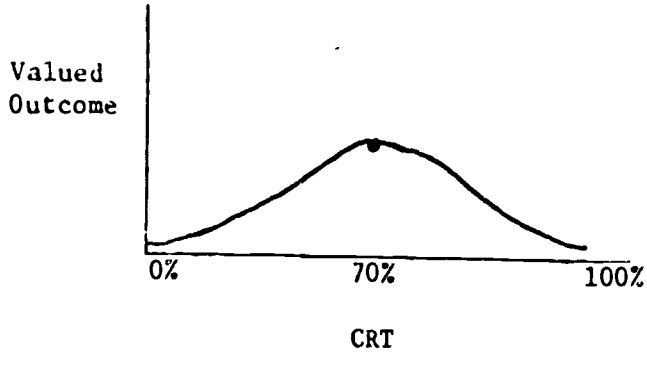
Some comments on Block's study are in line. One, the results lack generalizability. The 95% and 85% levels, which maximize the cognitive and affective measures respectively, are likely to change with the subject matter. Two, as pointed out by Glass (1978a), the method of maximizing a valued outcome assumes that there is a distinct point or

criterion score on the CRT that maximizes the outcome. What if the curve relating performance on the CRT is monotonically increasing, so that 100% performance on the CRT maximizes the valued outcome? In fact, it is more likely to be the case that the graph is monotonically increasing than the case where the graph increases and decreases. For example:

1. Monotonically increasing graph (Problem situation)




2. Ideal situation




(Reproduced from Glass, 1978a, permission for reproduction pending.)

Thus, it can be seen that unless the graph increases and then decreases, a 100% performance standard will be optimal. This standard is of limited use because it is not realistic to expect all students to attain that level.

Third, Block discusses that if there are multiple criteria to be maximized as valued outcomes, then some model for combining criteria with relevant weights needs to be developed. He does not offer any procedures for doing so however, and he looks at the effects of the performance standards on each of the 6 criteria separately. It should be noted that multiple criteria is a way around the problem discussed above (Glass, 1978a). For instance, if one of the outcomes has a monotonically increasing relationship with the test scores and the other a monotonically decreasing relationship, then the composite should have a peak value at a point other than 0% or 100%. While this would seem to solve the problem, another problem is only further exacerbated; what weights should be assigned to the valued outcomes to form the composite? These procedures have not yet been developed, and further, they are likely to be situation specific.

## 6.8 Combination Methods

### 6.8.1 Judgmental-Empirical

Zieky and Livingston (1977), and more recently, Popham (1978), have suggested two procedures that are based upon a combination of judgmental and empirical data. In addition, both Zieky and Livingston and Popham have

included an in-depth discussion of how to implement the procedures, something that has been lacking with many other procedures. The two procedures presented by Zieky and Livingston, the Borderline-Group and Contrasting-Groups methods, are procedurally similar. They differ in the sample of students on which performance data is collected. Further, while judgments are required, the judgments necessary are on students; not on items, as are many of the other judgmental methods (Nedelsky, Angoff, Ebel, etc.). Zieky and Livingston make the case that judging individuals is likely to be a more familiar task than judging items. Teachers are the logical choice as judges, and for them, the assessment of individuals is commonplace.

i. Borderline—Group Method

This method requires that judges first define what they would envision as minimally acceptable performance on the content area being assessed. The judges are then asked to submit a list of students (about 100 students) whose performances are so close to the borderline between acceptable and unacceptable that they can't be classified into either group. The test is thus administered to this group, and the median test score for the group is taken as the standard.

ii. Contrasting-Group Method

Once judges have defined minimally acceptable performance for the subject area being assessed, the judges are asked to identify those

students they are sure are either definite masters or non-masters of the skills measured by the test. Zieky and Livingston suggest 100 students in the smaller group in order to assure stable results. The test score distributions for the two groups are then plotted and the point of intersection is taken as the initial standard. This is exactly the same as the graphical procedure suggested by Berk, and presented in section 6.7.1. Zieky and Livingston then suggest adjusting the standard up or down to reduce "false masters" (students identified as masters by the test, but who have not adequately mastered the objectives) or "false non-masters" (students identified as non-masters by the test, but who have adequately mastered the objectives). The direction to move the cut-off score depends on the relative seriousness of the two types of errors.

iii. Suggestions

These methods, particularly the Contrasting-Groups Method, are very similar to the procedure suggested by Berk. Instead of actually forming instructed and uninstructed groups, however, as suggested by Berk, the Contrasting-Groups Method asks judges to form the groups. This judgmental procedure would seem more advantageous when the content being assessed has had a long instructional period (minimum competency testing is an example), or when there would be problems justifying the existence of an uninstructed group. Berk's method would be more useful for tests based on short instructional segments, most likely administered at the classroom level.

A comparison of the judgments involved in the two procedures indicates that the Contrasting-Groups Method would be the easier

method to justify using. It is a more reasonable task for teachers to identify "sure" masters and non-masters than it is for them to identify borderline students in the subject area being assessed. In sum, the Contrasting-Groups Method appears to us to be a most reasonable way of setting standards.

### 6.8.2 Bayesian Procedures

Novick and Lewis (1974) were the first to suggest that Bayesian procedures are useful for setting standards. Schoon, Gullion, and Ferrara (1978) have more recently discussed Bayesian procedures for setting standards. According to Schoon et al., Bayesian procedures allow the incorporation of:

1. A loss ratio, reflecting the severity of false-positive and false-negative decision errors,
2. prior information on the distribution of domain scores in the population of interest,
3. current information on an examinee's domain score, and
4. the degree of certainty that an examinee's domain score exceeds the cut-off score.

Of course, a cut-off score must first be set in order for the four factors to be incorporated. Thus, Bayesian procedures offer a way of augmenting the establishment of a cut-off score rather than a method for setting the cut-off score itself.

In sum, Bayesian procedures present a method for augmenting the setti of a cut-off score by utilizing available prior and collatoral information. The procedure also provides a posterior statement of degree of certainty about candidate's performance. Bayesian procedures do not, however, offer a method for setting a cut-score in the first place. Bayesian procedures have been included in this review because they do offer a method for combining judgmental and empirical data to arrive at a revised standard.

## 6.9 Some Procedural Steps in Standard Setting

In earlier sections of this unit, issues and many methods for standard-setting were discussed. In this section, procedures will be outlined for setting standards on criterion-referenced tests used for three different purposes. The purposes considered are:

1. Classroom testing
2. Basic skills testing for yearly promotion and high school graduation
3. Professional licensing and certification testing.

Classroom testing is emphasized since classroom teachers have fewer technical resources available to them than do the larger testing programs. Our ultimate objective is to provide a comprehensive set of practical guidelines for practitioners. At this time the guidelines are far from comprehensive; much research is needed to supply information necessary to construct thorough guidelines. We have suggested in places some of the questions that need to be answered.

Certain things are assumed: first, that in each case a set of objectives or competencies has been agreed upon, and that they are described via the use of domain specifications or some other equally appropriate method. Second, it is assumed that no fixed selection ratio exists (e.g., one might be fixed in effect by having resources to provide only a certain number of students with remedial work) since if it does there is no reason to set standards. Finally, we do not discuss the important and interesting political issues of who participates in and who controls the standard-setting process; we take as given that some such process exists and only address the issue of participation from the perspective of practicality.

6.9.1 Preliminary Considerations

Before any standard setting is undertaken for any purpose, an analysis of the decision-making context and of the resources available for the project should be done. The results of this analysis will determine how extensive and sophisticated the standard-setting procedure should be. Analysis of the decision-making context involves judging the importance of the decisions that are to be made using the test, the probable consequences of those decisions, and the costs of errors. Others have discussed using these same considerations in adjusting the final standard, but they may also be helpful in choosing a standard-setting method. Formal procedures for using this information are probably not necessary; a discussion of the issues by those directing the project should suffice. Some issues to consider would include (1) the number of people directly and indirectly affected by the decisions to be based on the test; (2) possible educational, psychological, financial, social and other consequences of the decisions; and (3) the duration of the consequences.

The next step should be a consideration of the resources available for the standard setting. Resources include money, materials, clock time, personnel time and expertise. How much of the total amount of available resources will be dedicated to the standard setting will depend upon the results of the prior discussion of decision context. The final decision as to the resources to be invested will determine how large and technically sophisticated the standard-setting enterprise may be.

A great deal of information needs to be collected on the actual expenditures of various resources that have been required to carry out

standard setting by different methods in different contexts. Actual time and money data would be invaluable to practitioners in choosing a method for their own situation. In the following discussion procedural steps in increasing order of expense and complexity will be offered but real data on these factors is lacking and is a pressing need.

### 6.9.2 Classroom Testing

The classroom teacher is most likely to use criterion-referenced tests for diagnostic purposes, that is for determining whether a student has mastered an area or needs further work in it. This would seem to be the most common situation calling for the setting of standards. Here the teacher must decide what level of test performance constitutes "mastery." In the same testing context the teacher may set additional performance standards, above and/or below the minimal level, for the awarding of grades on the material.

Typically the classroom teacher works alone, or at most with one or more other teachers of the same grade. It is also quite often the case that a classroom exam is used only once. In these situations methods based only on judgment of test content may be the only ones practicable. The methods developed by Ebel, Nedelsky and Angoff would be appropriate here, and the details of each of them have been discussed in an earlier section, so we will not re-iterate procedural steps here.

When available resources permit involving more people in the standard setting, parents and other community members might be enlisted, or a group of teachers of one grade from an entire school district might collaborate in setting standards. Again, if resources permit, data on group performance on individual items may be tabulated and considered in setting the standards on subsequent tests, or if tests are retained from year to year, the

performance data from the previous year might be used. Of course, this can also be done by teachers working alone. The following is a list of steps, some of which could be omitted if resources were limited, for involving parents of students in a particular class in setting standards for classroom tests over units of instruction. The method borrows heavily from Jaeger (1978). (It is assumed that the objectives have been identified and the teacher (or teachers) has prepared domain specifications):

1. At the beginning of the school year, a letter is sent to parents explaining the project and inviting them to a meeting where more information will be given.

2. At the meeting parents are given copies of domain specifications for the first test, along with example items. They are asked to indicate for each objective a percentage of items, which answered correctly would demonstrate the student had mastered the material adequately. At this meeting they should be encouraged to discuss the task and ask any questions they might have about it.

Instructions accompanying the standard-setting task should indicate to the parents how their judgment will be employed (for example, averaged with the percentages indicated by every other parent, and the resulting standard applied to every child in that class or grade). We have suggested for reasons of test security that the parents base their judgments on domain specifications rather than on actual test items; if test forms from previous years are available and thought to be parallel to the new exam, it may be easier for parents to make their judgments as a percentage correct of items on the parallel test.

3. The teacher constructs the criterion-referenced test from the domain specifications before looking at the parents' standards.

4. Class performance data is tabulated after the test is administered.

5. Parent judgment for the second test (or set of tests) is solicited by mail. The mailing packet includes: domain instructions (duplicating those given at the earlier meeting), and performance data from the first test (number of students achieving each set standard).

Instructions would also stress that judgments were to be based primarily on domain specifications and only secondarily on performance data.

6. Step 5 is repeated during the year whenever a competency-type test is to be given.

Alternatively, this procedure might be reserved for those instructional units judged to cover basic, required objectives for that grade; parents' instructions would then identify the tested materials as such.

7. The teacher keeps files for each test, including the domain specifications, parent judgment forms, actual exam and performance data.

8. Periodic meetings can be held to review the instructions and to discuss the procedure and its results.

Such discussions may lead to parents questioning the performance of students, and is likely to provoke query into both the teacher's methods and his/her subject matter. Teachers should be prepared for this; it may lead to parents wanting greater involvement in determining other aspects of their children's schooling, a desire one hopes can be creatively and constructively used.

Other variants on this procedure can include appointing a small committee of parents, possibly working with several teachers, instead of an open parents group. A parent-objective (matrix) sampling strategy could be employed to reduce the number of judgments required of each parent.

Another procedure for setting standards with criterion-referenced tests in instructional settings was offered by Hambleton (1978). According to Hambleton, "[His] is not a 'validated list' of guidelines. It is a list of practical guidelines I have evolved over the years through my work with numerous school districts." His eleven step list of guidelines is as follows:

1. The determination of cut-off scores should be done by several groups working together. These groups include teachers, parents, curriculum specialists, school administrators, and (if the tests are at the high school level) students. The number from each group will depend upon the importance of the tests under consideration and the number of domain specifications. At a minimum, I like to have enough individuals to form at least two teams of reviewers. This way I can compare their results on at least a few domain specifications to determine the consistency of judgments in the two groups. When sufficient time is available I prefer to obtain two independent judgments of each cut-off score.

2. I usually introduce either the Ebel method or the Nedelsky method. Following training on one of the methods, I have the groups work through several practice examples. Differences between groups are discussed and problems are clarified.

3. The domain specifications (or usually, but less appropriate, the objectives) are introduced and discussed with the judges.

4. I try to set up a schedule so that roughly equal amounts of time are allotted to a consideration of each domain specification. If some domain specifications are more complex or important I usually assign them more time.

5. I make sure that the judges are aware of how the tests will be used and with what groups of students.

6. If there exist any relationships among the domain specifications (or objectives) the information is noted. For example, if a particular objective is a prerequisite to several others it may be desirable to set a higher cut-off score than might otherwise be set.

7. Whenever possible I try to have two or more groups determine the cut-off scores. Consistency of their ratings can be studied, and when necessary, differences can be studied, and a consensus decision reached.

8. If some past test performance data are available, it can be used to make some modifications to the cut-off scores. On some occasions, instead of modifying cut-off scores, decisions can be made to spend more time in instruction to try and improve test performance. If past group performance on an objective is substantially better than the cut-off score, less time may be allocated to teaching the particular objective.

9. As test data become available, percentage of "masters" and "non-masters" on each objective should be studied. If performance on some objectives appears to be "out of line," an explanation can be sought by a consideration of the test items (perhaps the test items are invalid), the level of the cut-off score, variation in test performance across classes, a consideration of the amount of instructional time allotted to the objective and so on.

10. Whenever possible I try to compare the mastery status of uninstructed and instructed groups of examinees. Instructed groups ought to include mainly "master" students. The uninstructed groups should include mainly the "non-masters." If many students are being misclassified, a more valid cut-off score can sometimes be obtained by moving it (for example, see Berk, 1976).

11. It is necessary to re-review cut-off scores occasionally. Curriculum priorities change and so do instructional methods. These shifts should be reflected in the cut-off scores that are used.

There are many important questions needing to be researched. These techniques have apparently been used very little (there is certainly much more literature on how to set standards than on what happens when one does); we need to know the effects of involving different groups of people in the standard-setting (especially parents as opposed to others), of the number of people involved, the information and instructions provided and the frequency of standard setting. How do these factors effect the levels set, the public acceptability of the chosen standard, and are the procedures cost-effective?

### 6.9.3 Basic Skills Testing for Annual Promotion and High School Graduation

These are clearly areas where greater importance is attached to the consequences of testing and, hence, more resources will be allocated than for classroom testing. The discussion is limited here to testing of "minimal" competencies, not intending that the procedures be applied to the total curriculum. Further, we are not discussing the "life skill" or "survival" competencies; in setting standards for these skills it is necessary to consider performance on criterion measures of life success. We feel that this undertaking is beyond the capabilities of educational and measurement practice. It will be difficult enough to decide upon and assess "minimal" skills. For these skills, since no external criterion measures can be said to exist, the appropriate performance data to consider in standard setting are scores on the actual tests (or items). We agree with those (e.g., Jaeger, 1978; Linn, 1978; Shepard, 1976) who hold that performance data should be considered along with test content to inform the setting of standards. While from an idealistic point of view it would be desirable to set standards with reference only to the content of a domain, in reality the degree of skill in test construction required for the pure-content approach is probably beyond human attainment. In order to avoid unpleasant shocks it would seem good practice to examine test performance data; the other benefit of so doing is that feedback is received on our content-based judgments and may thus refine our skills.

Jaeger (1978) has provided an excellent guide to implementing a procedure involving representative groups affected by standards set for high school graduation. The method was discussed earlier, but a brief

review at this point seems useful. In general terms, it is an iterative procedure for soliciting item-by-item judgments from groups of judges. Information fed back to the judges at each iteration includes (a) group performance on each test item in a pilot administration, (b) the percentage of students who would have passed given several different standards, and (c) a distribution of the standards suggested by the judges in the group. The median passing score for each type of judge is computed, and the lowest of the medians taken as the standard.

The principal attraction of plans such as Jaeger's and the one outlined in Section 6.9.2, which is based on Jaeger's, is their political viability. By involving a broad cross-section of constituents in the setting of the standard, one increases the acceptability of that standard. However, no actual control or very significant influence over the educational process is transferred to the constituency; the objectives and the test, after all, are presented to them as givens, and their contribution in setting the standard is really quite limited. Moreover, the consensus method, while probably not harmful, may not produce results that make any pedagogical sense. Where obtaining popular support is not a critical problem, educators may prefer to rely upon the judgments of subject-matter and measurement "experts" to set standards. This may produce a more coherent, if less universally-accepted, result. Such a procedure could be implemented as follows (the steps would be executed for <u>each</u> subject matter area by content experts working with measurement experts):

1. Categorize the educational objectives or competencies as being of the knowledge/information type or of the rule-learning type (this distinction corresponds to Meskauska's (1976) continuum vs. state mastery models).

In the first case it makes sense to speak of a domain score, and to sample randomly from the domain to estimate that score. In the second, since

learning is presumed to be all-or-none, sampling considerations are not relevant, but construction of a few test items that accurately reflect the ability is critically important. Objectives domains of the first type reflect Ebel's (1978) notion of the purpose of competency certification tests as being efficient and accurate indicators of the level of achievement in a broad domain, rather than lists of specific competencies attained.

2. For objectives or competencies of the first type, construct tests with the aid of domain specifications, items matched to the domain specifications, and a suitable item sampling plan.

3. Ebel's standard-setting method (or one of the other content-focused methods) may then be used to set the standard for these parts of the test. To use Ebel's method the items from all of the knowledge/information (or continuum) domains would be considered together. (Table 6.9.3 provides a comparison of six possible methods.)

4. Pooling the judgments of all the experts may present a problem. Simply averaging the ratings given to each item (on relevance and difficulty) and/or the standards assigned to each category, will probably not give a very meaningful result. Ideally, the experts will go through a series of iterations in which they compare their independent judgments (first of the item categorization and next of the standards they assigned to each category), note discrepancies, discuss the rationale for each judgment, possibly decide upon revisions in the test (this will direct the procedure back to Step 2, to ensure that any revisions do not distort the test's domain representativeness), and/or p rsuade each other to change their judgments. Unanimity might be required in order to proceed from this step.

5. For those objectives or competencies classified as being of the "State" variety, smaller sets of items are required since the domains are more homogeneous, but item construction must be, if anything, more painstaking. Ideally, experimental evidence would be garnered to show that item performance truly reflected the target construct.

6. Standards on these State-type objectives can be adjusted back from 100% using Emrick's (1971) technique if the probabilities of Type 1 and Type 2 classification errors can be estimated. Similarly, domain scores can be adjusted by a Bayesian procedure (e.g., Hambleton & Novick, 1973) to compensate for relative losses associated with the classification errors.

Table 6.9.3

A Comparison of Several Standard Setting Methods

| Question | Judgmental | | | | | | Combination | |
|---|---|---|---|---|---|---|---|---|
| | Nedelsky | Modified Nedelsky | Angoff | Modified Angoff | Ebel | Jaeger | Contrasting Groups | Borderline Groups |
| 1. Is a definition of the minimally competent individual necessary? | Yes | Yes | Yes | Yes | Yes | No | No | Yes |
| 2. What is the nature of the rating task—or items, or individuals? | Items | Items | Items | Items | Items | Items | Individuals | Individuals |
| 3. Are examinee data needed? | No | No | No | No | No | No | Yes | Yes |
| 4. Do judges have access to the items? | Yes | Yes | Yes | Yes | Yes | Yes | Usually, but don't need to | Usually |
| 5. Are the judgments made in a group setting or individual setting? | Both | Both | Both | Both | Both | Both | Individual | Individual |

When the tests are used for yearly promotions, students' performance in the next grade can be used as a criterion in order to estimate the probabilities of classification errors.

Research is needed on ways of pooling the judgments of several individuals, and of incorporating performance data in primarily content-based judgments.

### 6.9.4 Professional Licensing/Certification Testing

Tests for licensing and certification differ from the others discussed here in having an external criterion, job performance, which the tests should predict. In addition, these tests are subject to governmental regulations and court rulings on the adequacy with which they reflect requisite job skills (and nothing more). Recent court decisions affirm that content validation of a test against the domain of entry-level job skills is sufficient to demonstrate that the test itself is fair. However, any standard used must also bear a rational relationship to job performance.

One method that will probably be acceptable in the courts is to base the standard on experts' judgments of the importance of each tested item to adequate job performance; that is, to use one of the content-oriented methods to determine a percent correct for passing. The pooled judgments of a large number of expert practitioners would be desirable.

Data on test performance would not be particularly useful in this situation since there is usually not any pre-existing knowledge or belief about the distribution of job-preparedness in the population. Empirical data on criterion (job) performance would be useful were it not for the pervasive selectivity of professions; to use criterion

performance properly in establishing optimal passing scores requires an unselected population of job-holders. For these reasons, content-oriented procedures for setting standards are probably the most viable procedures in licensing and certification.

Nassif, P. M. Standard setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of NCME, Toronto, 1978.

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. Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.

Popham, W. J. Setting performance standards. Los Angeles: Instructional Objectives Exchange, 1978.

Roudabush, G. E. Models for a beginning theory of criterion-referenced tests. Paper presented at the annual meeting of NCME, Chicago, 1974.

Schoon, C. G., Gullion, C. M., & Ferrara, P. Credentialing examinations, Bayesian statistics, and the determination of passing points. Paper presented at the annual meeting of APA, Toronto, 1978.

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

Torshen, K. P. The mastery approach to competency-based education. Ne York: Academic Press, 1977.

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

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

the standard and with the highest discrimination indices are selected for the test. Whether judges can reliably set standards from only domain specifications and some sample test items is unknown. Also, it is not known if standards set by these two different methods will produce different results. This is one of those situations where similar results across two methods would be highly desirable.

6.11 References

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, D.C.: 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.

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

Burton, N. Societal standards. Journal of Educational Measurement, 1978, 15, 263-271.

Conaway, L. E. Dissonant comments: Setting performance standards based on limited research. Florida Journal of Educational Research, 1976, 18, 35-36.

Conaway, L. E. Setting standards in competency-based education: Some current practices and concerns. Paper presented at the annual meeting of NCME, New York, 1977.

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

Ebel, R. L. The case for minimum competency testing. Phi Delta Kappan, April, 1978, 546-549.

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.

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

Glass, G. V. Minimum competence and incompetence in Florida. Phi Delta Kappan, 1978, 59, No. 9 (May), 602-605. (b)

Hambleton, R. K. On the use of cut-off 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.(Approx. Title) Berkeley, CA: McCutchan Publishing Co., 1979.

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

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.

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

Jaeger, R. M. Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 1976, 18, 22-27.

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

Klausmeier, H. J., Rossmiller, R. A., & Saily, M. Individually guided elementary education. New York: Academic Press, 1977.

Kriewall, T. E. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of AERA, Chicago, 1972.

Livingston, S. A. A utility-based approach to the evaluation of pass/fail testing decision procedures. Report No. COPA-75-01. Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service, 1975.

Livingston, S. A. Choosing minimum passing scores by stochastic approximation techniques. Report No. COPA-76-02. Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service, 1976.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

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

6.10 Summary

In this unit, a number of viable methods for setting standards were introduced. If you wish to view the test by itself and not in relationship to other variables, either Angoff's method or Nedelsky's method appears to be useful. If empirical data is available, Berk's method or the Constrasting Groups method seems especially useful. We have also discussed other methods, of a more complex nature, that are suitable for setting criterion-referenced standards. Our preference for the methods mentioned above stems from the fact that they are simple to implement, and appear to produce defensible results when applied correctly. In the final section of the paper, some proposed sets of procedures for standard setting with respect to three important uses of criterion-referenced tests were outlined. However, considerably more research must be done before these procedures can be recommended for wide-scale use.

We will conclude this unit with a brief discussion of a very important problem. Suppose a set of test items have been selected. If so, it is then possible to set standards via either judgmental or empirical methods (or both). However, if a standard can be set via reference to well-defined domain specifications, and sample test items, tests which will optimally discriminate (i.e., reduce the number of misclassifications) in the region of a standard can be constructed. This is done by selecting test items which "discriminate" in the region of the standard. Test items are piloted on samples of examinees similar to those who will eventually be administered the tests to determine item difficulty levels and discrimination indices. Items with p values near

## Additional References

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 two cutting score procedures using generalizability theory. ACT Technical Bulletin No. 33. Iowa City, Iowa: American College Testing Program, 1979.

Eignor, D. R. Psychometric and methodological contributions to criterion-referenced testing technology. Unpublished doctoral dissertation, University of Massachusetts, Amherst, 1979.

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

Levin, H. M. Educational performance standards: Image or substance? Journal of Educational Measurement, 1978, 15, 309-319.

Linn, R. L. Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 1978, 15, 301-308.

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

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