DOCUMENT RESUME

ED 206 685 TM 810 583

AUTHOR Mehrens, William A.
TITLE Setting Standards for Minimum Competency Tests.
PUB DATE 24 Feb 81
NOTE 35p.; Revision of a speech presented at the Michigan School Testing Conference (Ann Arbor, MI, February 24, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; *Cutting Scores; Elementary Secondary Education; *Minimum Competency Testing; *Scoring Formulas; *Standards
IDENTIFIERS Angoff Methods; Compromise Model (Hofstee); Ebel Method; Empiricism; Jaeger Method; Nedelsky Method

ABSTRACT

Some general questions about minimum competency tests are discussed, and various methods of setting standards are reviewed with major attention devoted to those methods used for dichotomizing a continuum. Methods reviewed under the heading of Absolute Judgments of Test Content include Nedelsky's, Angoff's, Ebel's, and Jaeger's. These methods are compared and a preference for Jaeger's approach is stated. Under Standards Based on Judgments about Groups, the Zieky and Livingston contrasting group and borderline group methods are discussed. The approaches proposed by Berk and Block are briefly discussed as Empirical Methods for Discovering Standards. A summary statement lists some "DO NOT'S" and "DO'S" for setting cutting scores. (Author/GK)

Setting Standards for Minimum Competency Tests*

William A. Mehrens

Michigan State University

*Revision of a speech given at the Michigan School Testing Conference, Ann Arbor, Michigan, February 24, 1981.

ABSTRACT

First some general questions about minimum competency tests are discussed. Then various methods of setting standards are reviewed with major attention devoted to those methods used for dichotomizing a continuum. Methods reviewed under the heading of Absolute Judgments of Test Content include Nedelsky's, Angoff's, Ebel's, and Jaeger's. These methods are compared and a preference for Jaeger's approach is stated. Under Standards based on Judgments about Groups, the Zieky and Livingston contrasting group and borderline group methods are discussed. The approaches proposed by Berk and Block are briefly discussed as Empirical Methods for Discovering Standards. A summary statement lists some "DO NOT'S" and "DO'S" for setting cutting scores.

1. Introduction

A. History of Minimum Competency Tests

As many others have pointed out before me (e.g., Ebel, 1978), minimum competency testing has been around for a long time. A very early minimal competency exam was when the Gilead Guards challenged the fugitives from Ephriam who tried to cross the Jordon river.

"Are you a member of the tribe of Ephriam?" they asked. If the man replied that he was not, then they demanded, "Say Shibboleth." But if he couldn't pronounce the "sh" and said Sibboleth instead of Shibboleth he was dragged away and killed. So forty-two thousand people of Ephriam died there at that time. (Judges 12:5-6, The Living Bible).

Nothing is reported concerning the debates that may have gone on among the guards regarding what competencies to measure, how to measure them, when to measure, how to set the minimum standard, or indeed what should be done with the incompetent. We do not know the ratio of false positives to false negatives or the relative costs of the two types of errors. We do know that a very minimal competency exam was given and that forty-two thousand people failed - with no chance of a retake. And some people in Michigan think they have it bad!

But there have been other, less drastic competency exams - for example those for certifying or licensing professionals and those for obtaining a driver's license.

If not a new concept, why so much fuss? Never before have state and local agencies been so active in setting the minimum competency standards for elementary and secondary students. At least 35 states have taken some such type of action, and it has been reported (Pipho, 1978)

that all the remaining states either have legislation pending or legislative or state board studies under way.

B. General questions about minimum competency tests

Over the past several years a multitude of questions have been raised about minimum competency testing. For example: (1) why have them at all, (2) what competencies should be measured, (3) how should we measure them, (4) when should we measure the competencies, (5) who should set the minimum standard, (6) how should the mimimum standard be determined, (7) should there be one minimum or many, and (8) what should be done with the incompetent? These questions are all related. The answer given for one has implications for the answers for the others. Thus, although my charge today is to discuss question 6,-how should the standard be determined? - it seems advisable to briefly mention my views on the answers to the other questions. Further details regarding my views of all these questions can be found in Mehrens (1979).

(1) Why have standards at all?

Why the big push for minimum competency tests with specified standards? Many individuals believe the evidence suggests that the quality of our children's education is deteriorating and that minimum competency testing will improve educational quality (or reverse any deterioration). Both points are debatable. I believe the first - some of you may not. The second point is one where I would prefer to reserve judgment but, as mentioned, there is some supportive evidence reported in the literature.

Of course there are many perceived costs as well as perceived benefits of minimum competency testing. Perkins (in Gorth and Perkins, 1979), has compiled the following two lists:

Perceived Costs of Minimum Competency Testing

- emphasis on the practical will lead to an erosion of liberal education
- causes less attention to be paid to difficult-to-measure learning outcomes
- promotes teaching to the test
- will be the "deathknell for the inquiry approach to education"
- oversimplifies issues of defining competencies and standards and of granting credentials to students
- promotes confusion as to the meaning of the high school diploma when competency definition is left to local districts
- fails to adequately consider community disagreement over the nature and difficulty of competencies
- will exclude more children from schools and further stigmatize underachievers
- will cause "minimums" to become "maximums," thus failing to provide enough instructional challenge in school
- may unfairly label students and cause more of the "less able" to be retained
- may cause an increase in dropouts, depending on the minimum that is set
- provides no recognition of the "average" student

- fails to provide alternatives that can "inspire" average students to excell in some areas
- ignores the special needs of gifted students, giving them less opportunity to be challenged and to expand their horizons
- may have adverse impact on a student's future career as a result of a witheld diploma
- may promote bias against racial, ethnic, and/or special needs groups
- places the burden of "failure" on the student
- causes educators to be held unfairly accountable
- intensifies the conflict for educators between humaneness and accountability
- increases the record-keeping burden for administrators
- does not assure that students will receive effective remediation
- does not assure that all of the perceived needs and benefits will be met and realized
- promotes the power of the state at the expense of local district autonomy
- can be costly, especially where implementation and remediation are concerned

Perceived Benefits of Minimum Competency Testing

- restores meaning to a high school diploma
- reestablishes public confidence in the schools
- impels us to face squarely the question of "what is a high school education?"
- sets meaningful standards for diploma award and grade promotion

- challenges the validity of using seat time and course credits as basis for certifying student accomplishments
- certifies that students have specific minimum competencies
- involves the public and local educators in defining educational standards and goals
- focuses the resources of a school district on a clear set of goals
- defines more precisely what skills must be taught and learned for students, parents, and teachers
- promotes carefully organized teaching and carefully designed sequential learning
- reemphasizes basic skills instruction
- helps promote competencies of life after school
- broadens educational alternatives and options
- motivates students to master basic reading, mathematics, and writing skills
- stimulates teachers and students to put forth their best efforts
- identifies students lacking basic skills at an early stage
- encourages that schools help those students who have the greatest educational need
- can bring about cohesiveness in teacher training
- encourages revision of courses to correct identified skill deficiencies
- can truly individualize instruction
- shifts priorities from process to product
- holds schools accountable for educational products
- furnishes information to the public about performance of educational institutions
- provides an opportunity to remedy the effects of discrimination by identifying learning problems early in the educational process

- provides an opportunity to remedy the effects of discrimination by identifying learning problems early in the educational process
- provides greater holding power for students in the senior year
- provides for easier allocation of resources

Shepard (1980) bypasses the cost/benefit debate to discuss the three primary uses of competency test scores: pupil diagnosis, pupil certification, and program evaluation. The various methods used to set standards are differentially appropriate for the various intended uses.

(2) What competencies

The answer to the question of what competencies should be measured in a minimaal competency program is related directly to the purposes of the test, i.e., what inferences we wish to make about a person who "passes", and much less directly about the "purposes of the school." Many people apparently do not make enough of this distinction.

Although there exists a reasonable consensus about desirable adult characteristics, there is considerable diversity of opinion about their relative importance and the role of the school in promoting those characteristics. Some people maintain that good citizenship or healthy self-concepts are more important in life than reading skills. Others assert just the opposite. And some who believe the former do not believe it is the primary purpose of the school to promote those characteristics. I suspect we will

never reach agreement on what characteristics we "need" in our society and on the role of the school in teaching, establishing, or nurturing those characteristics. That does not dismay me, nor do I believe it should deter us from determining general content for a minimal competency test. No test can be designed to assess the degree to which all the purposes of education have been achieved or even to assess whether students have achieved a level of minimal competency in all areas.

Surely no one would infer that all purposes of education have been achieved if students pass a minimum competency test. What will people infer and/or what do we want people to be able to reasonably infer from a passing score on a minimum competency test?

Would any reasonable citizen infer - or would we want them to infer - that a passing score means the person has "survival skills" for life? Life is very varied, and so are the skills needed to survive. I cannot believe the populace is so unrealistic or naive as to think in such grandiose terms. Schools do not and canno teach all survival skills. Such skills cannot even be adequately enumerated (or defined), and thus they cannot be adequately measured. Since we do not want any "survival skills" inference to be drawn from a test, we should not build a test to measure such defined competencies.

But if we measure only basic skills (applied to life settings), won't other areas of school suffer? I do not htink so, Remember, there is a distinction between the purpose of school and the purposes of a minimal competency test. The purpose of the latter can never be to assess all the objectives of school. We all know that. Of course not all skills are basic and we do not want mi imums to become maximums. Few would be happy to see high school graduates who lacked maturity, self-discipline, and some understanding of their own value systems. But if we keep in mind the limitations of the inferences to be drawn from passing (or failing) a minimum competency

test, such limited testing should not have deleterious effects.

We should not assume that minimal competency standards can do much at all to define the goals and objectives of education. They only set a lower limit of acceptable standards in certain basic skill areas. This certainly suggests that passing the minimal competency test should not be the only requirement for high school graduation. Other graduation requirements could assure breadth in other areas. In specifying the domain of basic skills, we need to keep in mind the relationship between the tested domain and what is taught in school. We should not be testing content that is not taught. On the other hand, we should not attempt to randomly sample all that is taught. The tested domain must be a subset of materials taught in the curriculum. The domain must be defined precisely enough to rule out relatively unimportant specific bits of factual knowledge as well as processes so abstract they appear to measure general intelligence. There should be evidence not only that the material tested is actually taught (i.e., presented) but that almost all students are capable of mastering the materials.

3. How to Measure

There are a variety of possible meanings to the question "How to Measure." How to sample, how to administer the measures, or how to build the measuring instruments are all possible meanings. Most people who have spoken to this question have addressed the latter point - usually with respect to the type of measuring instrument. Obviously, the choice depends somewhat on what competencies one wishes to measure. Remember, I have voted for basic skills. I believe for

such competencies we can get along reasonably well with what we have traditionally called objective paper-and-pencil tests, but the answer must depend on the specific domain definitions of the competencies.

4. When to Measure

The answer to the question "When to Measure" (like the answer to every other question) depends on the purpose or purposes of testing. Of course the primary reason for minimal competency testing is to identify students who have not achieved the minimum. But, identify for what purpose? To help the students identified through remediation program? To motivate students through "fear of failure?" To make a high school diploma more meaningful? Let me assume the answer is yes to all questions.

First, let me suggest that there should be periodic but not every year testing. I believe minimum competency programs will be more cost-effective if tests are given approximately three times during the K-12 portion of a student's schooling, for example in grades 4, 7 and 10. Teachers, of course, gather almost continuous data. They often have already identified those students achieving inadequately. The formal tests supplement the teachers' measures and confirm or disconfirm previous judgments. I believe that this formal identification is useful. Tests are credible instruments, help motivate students (and teachers), and help assign a minimal competency meaning to a diploma or certificate. I would like to stress, however, that while I favor tests it is NOT because I believe that teachers' judgments without them would be grossly inaccurate.

I am opposed to every grade testing for minimal competencies because it is not cost-effective.(I am not opposed to every grade testing with a more general achievement measure.)Only a very few

students, we hope, will be identified as not achieving at a minimal level, and at any rate those identified in fourth grade would very likely overlap considerably with those in third or fifth grade. Further, annual testing may result in grade by grade promotion decisions based on test results. In spite of the generally favorable press for this approach in Greensville County, VA, I remain somewhat skeptical about such a plan.

I selected early grades because I believe the evidence shows that remediation is more effective when begun early. I believe we need high school testing for two reasons: (1) Not all students will "make it" by seventh grade and (2) those who do many need to recheck their skills. Forgetting does occur between grades seven and ten - especially if the material is not part of the curriculum in the intervening period.

Finally, let me stress that if minimal competency tests are used for high school certification or graduation, there must be opportunities for students who have not passed to retake the exams. Further, no test should be used for such a purpose the first year it is given. To be fair to students there should be a phase-in period.

5. Who Sets the Minimums?

Obviously, the minimums must be determined by those who have the authority to do so. This is an agency such as a state board of education or a local school board. It is more difficult to decide who should represent this agency. Of course all constitutents should be involved, but I firmly believe that measurement experts need to be involved as well. Although setting the minimum is arbitrary, measurement experts can have some useful suggestions. These should become obvious as we discuss the various methods of setting the standards.

6. One Minimum or Many?

The answer, again, depends on purpose. For example, do we wish to categorize or diagnose?

Assume for the moment that two basic skill areas have been identified. If we are truly concerned with minimum performance in each area, then we can not use a single total score or any type of compensatory scaling model across areas. Multiple cut-off scores are needed -- at least one for each basic skill area.

But what about the subskills (objectives) within an area? Will a compensatory model work there, or do we need multiple cut offs? If the latter, would we require a "pass" on every subskill objective or only on a certain percentage of them? Of course the answer, and the importance of the answer, will depend upon the covariance structure of the item scores within and across objectives and objective scores within the total test. If the test covers heterogeneous competencies - each important - then a single cut off score would not be too meaningful. Setting separate standards for each objective, however, results in another problem - that of the reliability of the scores. The empirical evidence I am aware of suggests that the covariances across objectives are sufficiently high so that one can defend the use of a total score within the broader basic skill. Of course if it seemed useful, one could report out objective by objective and still have only one total cut off score for the categorization decision. The usefulness of the objective by objective information would be dependent upon the reliabilities of each objective score as well as the reliabilities of the difference scores.

Even if Scores are reported objective by objective, the total score should be based on the total number of items correct, not the total number of objectives passed. Unless differential weights are used for the objectives, the former method is the more reliable. To dichotomize objective scores before combining them results in the loss of information.

Later in this paper I will address in detail one of the general questions raised at the beginning of this section - how should the minimum standard be determined. First let me mention some general points which anyone involved in setting cutting scores should consider.

II. General points to keep in mind when setting the cutting score

Before choosing any particular method of setting cutting scores, the tester should consider the legal defensibility of the procedure, the ease and cost of implementing the procedure, public acceptance of both the procedure and the cutting score, its psychometric characteristics, and political considerations. (Nassif, in Gorth & Perkins, 1979).

Discussing the legal defensibility of decisions made on the basis of minimum competency scores would require far more time than is available today. Certainly legal questions concerning the authority to make the decision and the reliability, validity, and potential bias of the measuring device could be raised. But today the focus is not on these issues but rather on whether the cutting score is set appropriately. "Appropr'ateness" in licensing or certifying decisions in industrial psychology would probably require empirical and/or logical relationships between the cutting score and the minimum ability required to do the job. In education "appropriately" probably means

following one of the established procedures.

Obviously some of the procedures to be discussed are much more complicated than others. The choice of method is affected by the availability of money, time, and technical expertise.

Also, the method chosen should provide reliable, valid, and unbiased results. One should pay attention to the false positive and false negative rates and the relative costs of those two types of errors. The desirability of public acceptance is increasingly important. Three primary factors influencing public acceptance are the ease with which the process can be understood, the involvement of the public in the process used, and the proportion of the test takers who fail to reach the cut off score. Although political considerations will not be discussed here, anyone responsible for setting cutting scores should become aware of the political climate of the district or state and the implications of this climate for setting cutting scores.

III. Methods of Setting Standards

A. Introduction

First off I would like to admit, like others before me, that the actual choice of a minimum is arbitrary. Different methods of setting the minimum lead to different cutoff scores, and one cannot say in the abstract that one methods (or one cutoff score) is superior to another. Gene Glass makes the point as follows:

> "I have read the writings of those who claim the ability to make the determination of mastery or competency in statistical or psychological ways. They can't. At least, they cannot determine "criterion levels" or standards other than arbitrarily..... the language of performance standards is pseudoquantification, a

meaningless application of numbers to a question not prepared for quantitative analysis." (Glass, 1978a, p. 602)

So, admittedly, setting the standard is arbitrary. Further, it is politically and economically influenced. If the standards are too high and too many students fail, then there will surely be a public outcry about the quality of the schools and the unreasonableness of the standards. Further, if one is committed to remediation, the costs of remediation could be very high. If the standards are set too low then the program becomes meaningless, and if people become aware of the ridiculously low standards, they will again present an outcry about the quality of the schools. The standard setters will be damned either way.

Glass raises the question of whether a criterion-referenced testing procedure entailing mastery levels is appropriate. He answers in the negative stating that "nothing may be safer than an arbitrary something." (Glass, 1978b, p. 258)

Now I certainly admire Gene Glass as a person and I agree with much of what he has said in the two articles I have referenced in this section. And indeed, we might be "safer" with nothing rather than an arbitrary something. But let me for the moment take the other side.

There is no question but that we make categorical decisions in life. If some students graduate from high school and others do not, a categorical decision has been made whether or not one uses a minimal competency exam. Even if everyone graduates, it is still a categorical decision if the philosophical or practical possibility of failure exists. If one can conceptualize performance so poor the performer should not graduate, then theoretically a cutoff score exists. The

proponents of minimal competency exams seem to believe, at least philosophically, that there is a level of incompetence too low to tolerate, and that they ought to define that level so it is less abstract, less subjective, and perhaps a little less arbitrary than the way decisions are currently made.

The above is not an argument for using minimal competency test alone as graduation requirements. Nor is it an argument for using a dichotomous (as opposed to continuous) test score as one of the factors in that decision. What I am trying to make very clear is that ultimately - after combining data in some fashion - a dichotomous categorization exists: those who receive a diploma and those who do not. No matter what type of equation is used, linear or nonlinear, no matter what variables go into the equation, no matter what coefficients precede their values, the final decision is dichotomous and arbitrary. The argument against minimal competency exams can not be that they lead to an arbitrary decision unless one truly believes that all individuals - no matter what their level of performance - belong in the same category.

If someone has decided to set an observed minimal test score, how should it be done? Theoretically this is no problem. Decision theory spells out exactly how to proceed. First, determine the "true" mastery level cutting score and the cost of false positives and false negatives. Then some simple mathematics will show where to set the observed cutting score (or, more precisely, how to allocate individuals to mastery states) such that the total cost of errors will be minimized. Of course we do not know what values to give to "true" mastery level or to the cost of the false positives and false negatives!

Practically, there are many different ways that have been suggested

These are thoroughly discussed in readily available literature, and readers wishing a more thorough presentation should check Millman (1974), Glass (1978b), volume 15, #4 of Journal of Educational Measurement (Winter 1978), Hambleton (Ch. 4 in Berk, 1980), Nassif (Ch. 4 in Gorth & Perkins, 1979), and Shepard (1980).

The various authors mentioned above categorize the methods differently. For example, Hambleton talks about judgmental, empirical, and combination methods. Glass categorizes the methods as 1) performance of others, 2) counting backwards from 100%, 3) boot strapping on other criterion scores, 4) judging minimal competence, 5) decision-theoretic approaches, and 6) operations research methods. Shepard has two major categories: methods which assume mastery is an all-or-none state, and methods for dichotomizing a continuum. She provides five subcategories for the second class. In this presentation I will basically follow her outline but precede her categories with one which Nassif calls "administrative decision or consensus."

B. Administrative decision or consensus

Nassif points out that administrative methods can not be classified as using either judgmental or statistical assumptions because they have little structure. These methods, however, are very commonly employed. "Setting standards by administrative decision means simply that the cut off score is determined by one or more persons holding a position of authority: (Nassif, in Gorth & Perkins, 1978, p. 105). This may or may not be an informed decision, it may or may not be based on any data. It is an easy method to use and, if the person making the decision actually has the authority to do so, it permits some legal defense; but a good prosecuting attorney would

make this process seem pretty inadequate. It is not a method which will necessarily win public acceptance, although the person setting the score may be quite sensitive to public, financial, and political concerns. There is no reason to believe such a method will lead to a cutoff score with appropriate psychometric characteristics.

The consensus method is similar except that the decision is made by a group of people who either have or have been allocated the decision making power. If no specific method is used by the group this procedure has the same advantages and limitations of the administrator decision making approach. If some specific methodological procedure is followed, we would classify the procedure other than simply "consensus".

C. Methods which assume mastery is an all-or-none state (counting backwards from 100%)

Some standard setting models (state models) assume that mastery is an all or none affair - an examinee either has the skill or does not. If a person is a master he/she should be able to get all the items correct except for those missed due to measurement errors. Thus, the standard setting task involves a question of how much to adjust the 100% standard downward.

"Just how great a concession is to be made becomes distressingly arbitrary, with some allowing a 5% shortfall and others allowing 20% or more." (Glass, 1978b, p.244)

Advocates of such a procedure usually ignore the fact that items measuring a specific objective may very greatly in difficulty. Since I (and most others whose writings I have read) believe the all-or-none assumption is not very plausible, these methods will not be considered further.

D. Methods for Dichotomizing a continuum

In continuum models the characteristic being assessed is assumed to be continuous. The cut off score is chosen such that it is the least amount a person can score and still be considered a master. "All of the methods proposed to formalize the selection of this cut off point are decision strategies to help in thinking about what amount of knowledge should be required" (Shepard, 1980, p.451).

1. Absolute Judgments of Test Content

Criterion referenced testing typically results in absolute rather than relative interpretations. Thus, to many people it seems reasonable to simply inspect the test content and to decide what percentage of correct answers indicates mastery. We will briefly consider four such methods: Nedelsky, Ebel, Angoff, and Jaeger.

a) Nedelsky's approach

The Nedelsky (1954) approach is the oldest of the procedures and has been used considerably in the health professions which is the area for which the procedure was developed. It can only be used for multiple-choice questions with right answers. Basically, the Nedelsky procedure involves asking each of a set of judges to look at each item and identify the incorrect options that a minimally competent individual would know were wrong. Then, for each judge, the probability of a minimally competent student getting an item correct would be the reciprocal of the remaining number of responses, (e.g., if on 5 alternative item, a judge feels a minimally competent student could eliminate 2 options, than the probability of such a person getting the item correct is 1/3). The expected score on the

test for a minimally competent student would be the sum of the obtained reciprocals across all items. Of course not all judges will come up with the same score so the total set of minimally competent scores for the judges are averaged ($\bar{X}$). According to Nedelsky, the standard deviation of the judges' scores would be equal to the standard deviation of the scores of minimally competent students. Thus, this standard deviation (σ) could be multiplied by a constant K (decided by the judges, or test users) to regulate the percent of minimally competent who pass of fail. Thus the final cut off score is:

$$C.S. = \bar{X} + K\sigma$$

Assuming an underlying normal distribution if one wishes 50% of borderline examinees to fail one sets K = 0, if one wishes 84% to fail one set K = 1, if one wishes 16% to fail one sets K = -1, etc.

b. Angoff and modified Angoff approaches

The Angoff method is similar to Nedelsky's only the judges are not asked to delete options but just to estimate the probability that a minimally acceptable person would get each item right. The sum of the probabilities becomes the cut off score. ETS has simplified this procedure somewhat by providing a seven point scale on which percentages of minimally knowledgeable examinees who would get the items right are fixed (5,20,40,75, 90,95, Do Not Know) and asking judges to mark this scale.

c. Ebel's approach

In Ebel's (1972) approach the judges are asked to rate the items on the basis of relevance (4 levels) and difficulty (3 levels). These categories form a 4 x 3 grid. Each judge is

asked to assign each item to the proper cell in the grid and also, once that is done to assign to the items in each cell a percentage correct that the minimally qualified person should be able to answer. (This percentage may be agreed on by the judges via some process, or one could proceed with each judge's values and average at the final stage.) Then the number of questions in each cell is multiplied by the percentage to obtain a minimum number of questions per cell. These numbers are added across the 12 cells to get the total number of questions the minimally qualified person should be able to answer.

Example for one judge

	Difficulty Level						
	Easy		Medium		Hard		Summed
Relevance	# Items	% Correct	# Items	% Correct	# Items	% Correct	Number x %
Essential	40	100%	15	80%	10	30%	55
Important	5	90%	10	70%	10	20%	13.5
Acceptable	5	90%	5	40%	0	10%	6.5
Questionable	0	70%	0	50%	0	0%	0

Cutting Score = 75

d. Jaeger's approach

This approach is primarily judgmental but does use some normative information so others may place this process in some other category. In one specific example (Jaeger, 1978) 700 people were divided into 14 groups of 50 each. In each group everyone took the test and then answered two questions on each item:

1) Should every high school graduate be able to answer this item correctly?

2) If a student does not answer this item correctly, should s/he be denied a high school diploma?

After each judge finishes this they receive the overall results of the survey and their test performance. Then they are asked to review and revise their standards. Finally they are told the proportion of students who would have failed based on the recommended cut off score and asked to reconsider their ratings and make a final judgment regarding the necessity of passing each item on the test. Finally a median score is caluculated for each group and the cutting score is set at the lowest median cutting score given by the groups.

e. Comparison of above methods

There is no question but that different methods produce different curring scores (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Kleinke, 1980; Skakum & King, 1980). For example, in the Skakum & King (1980) study the Nedelsky method resulted in 23% failure rate and the Ebel method a 46% failure rate.

There is no compelling theoretical reason to prefer one of the above methods to any other. Most writers seem to prefer Angoff for its simplicity. I prefer Jaeger's approach in that it provides normative data but, as mentioned, it therefore maybe does not belong in this category of methods.

f. Problems and considerations of these methods

1. They do not agree with each other.

2. There is considerable disagreement among judges within method and the averages obscure this.

3. It is difficult to build a theoretical rationale for any of these models.
4. Such issues as what value to set for K in the Nedelsky approach and how many cells to use in the Ebel approach allow for considerable differences within variations of any one approach.
5. Standards are often set quite high under these approaches and thus many people fail.

2. Standards based on judgments about groups

As Shepard points out, judgments based on test content alone can result in standards that are obviously wrong. Sometimes individuals fail such tests when other evidence of their mastery is more compelling than the belief in the accuracy of the standard. In an attempt to avoid such situations some people advocate setting the standard by looking at the performance of individuals in an identified group.

a. Zieky & Livingston: Contrasting Groups

In this approach, judges (teachers perhaps) are asked to pick individuals that clearly belong to one of two groups of examinees (using available information <u>other than</u> the test): one group composed of individuals who are clearly masters and another composed of individuals who are clearly nonmasters. The test is then given to both groups, the distributions are plotted and an initial standard is set at the intersection point of the two plots. Then, if judgments are available about the relative costs of false positives and false negatives the cutting score can be raised or lowered to minimize the total cost of the misclassifications.

b. Koffler

Koffler (1980) uses a quadratic discriminant function to set the cutting score, otherwise the approach is the same as the contrasting Groups approach.

c. Zieky & Livingston: Borderline Group method

This method is similar to their Contrasting Groups method except the judges are to choose individuals who they believe are borderline with respect to minimal competency. This group is given the test and the standard is set at the median. (Of course, one could choose to pass some other percentage of minimally competent individuals. This would be analoagous to setting a K value in the Nedelsky approach). This approach is generally considered to be inferior to the Contrasting Groups approach because it is more difficult to identify an adequate sample of borderline examinees.

3. The Use of Norms

In some of the previous methods discussed, empirical data gathering techniques were used to help set standards but the standard was not based on a conscious decision to fail any given percent of the total set of individuals. To choose a cutoff score by a normative approach seems, to some, to be contradictory to the purpose of criterion referenced testing. But even Popham now admits we should norm our criterion referenced tests (Popham, 1976). As Shepard has pointed out (and others before her)".... it is only the first use of criterion-referenced tests, estimating domain scores, that can be accomplished without relative comparisons. Qualitative judgments about the excellence or adequacy of performance depend implicitly on how others did on the test. Expectations about

what a lawyer or high-school graduate should know are normative. If everyone could intuit the theory of relativity on their way to work, Einstein would not have been considered a genius" (Shepard, 1980, p. 456).

Certainly cut off scores set without any normative data can be very embarassing. Rentz (1980) tells how a Georgia teacher certification examination had a cutting score three standard errors below what was considered the very least one should know but when the test was given too few passed it, so a new cutting score was set consistent with a desired pass rate.

Whether or not one should use only a normative group and a desired pass/failure ratio is of course debatable. But leading writers now seem to agree that at least normative data could well be helpful to decision makers when used in conjunction with some other method.

4. Hofstee's Compromise Model

Hofstee (1980) has proposed a compromise model in which judges are asked to specify the following values.

1. The Maximum required percentage of mastery: K-max. This is the cut off score which would be satisfactorily high even if every student scored that high or higher.
2. The minimum acceptable percentage of mastery: K-mim. This is the cut off score which is as low as one would go even if no student attained that score.
3. The maximum acceptable percentage of failures: F=max
4. The minimum acceptable percentage of failures; F=min.

Hofstee then graphs the two (dimension-test score and percent passing) and uses a formula for arriving at a midpoint between F-min, K-max and

F-max, K-min.

5. Empirical methods for discovering standards

a. Berk's instructed and uninstructed groups

This approach is very similar to the Contrasting Group method of Zieky and Livingston. The distinction is that one does not use judgment to determine who goes in which group. Rather the two groups are determined as those who have been instructed and those who have not. As with the contrasting groups procedure one can either set the standard to minimize the total number of errors or one can differentially weigh the false positives and false negatives. Berk's procedure is most appropriate for instructional decision making. As Shepard (1980) has pointed out, this procedure will not work for high school minimum competency testing because a) one can not identify instructed and uninstructed groups and b) the assumption that the instructed group will be predominantly masters is not necessarily valid.

b. Block's educational consequences

(Glass' operations reasearch method)

In this method one attempts to set the cutting score to maximize future learning or other cognitive or affective criteria. The question is "What passing score maximizes educational benefits?" This method assumes there is some "functional relationship between performance on the test and level of performance on the criterion variable" (Shepard, 1980, p. 549). Actually, I know of only one study that has used this approach (Block, 1972). It, like the Berk method is appropriate only for instructional decisions

making, not for certification decisions. There are several problems in Block's approach (Glass, 1978b; Hambleton & Eignor, 1979) and it is not one I would recommend.

6. Empirical methods for adjusting standards

(Glass' decision-theoretic catagory)

The methods classified under this approach use a decision theory and attempt to set cutting scores to ensure a minimum cost of the errors. These methods are different from those that determine standards because they presume a standard already exists on an external criterion and the various methods translate this external standard into a cut off score on the test. This means that someone has already had to make some decision with respect to a standard on the criterion. Obviously if an external criterion does not exist, these approaches cannot be used. For that reason they are not likely to be useful in minimum competency testing programs for high school graduation since there is no standard of "adult success".

Since these approaches are not likely to be useful and since they are fairly technical we will not review them here. Those of you interested could check Huynh (1976), Livingston (1975), Novick and Lindley (1978) and Vander Linden and Mellenbergh (1977).

IV. How to choose a standard-setting method

A. Factors to consider

Generally in selecting a method, someone would keep in mind the points discussed earlier, such as legal defensibility, ease of implementation, financial factors and public acceptance. One should also consider the importance of the decision, the qualifications of the judges (since some methods require more knowledgeable judges than others), and the appropriateness of the method

for the type of decision (Hambleton, in Berk 1980).

B. Uses of Data

1. Pupil diagnosis

As Shepard (1980) pointed out, classroom passing scores are usually set informally because teachers do not have the knowledge or resources to use the more elaborate methods. Classroom errors in classification, moreover, are not so costly. The best advice to give teachers is to keep in mind the relative costs of advancing someone who should be retained versus retaining someone who should be promoted.

2. Pupil certification

Shepard has made such a good statement about this that I wish to quote her extensively:

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

the test) and their beliefs about how many individuals are qualified." (Shepard, 1980, p. 463)

The Angoff and Jaeger methods are generally considered the most practical approaches of judging test content. If qualified judges of people exist, the Zieky & Livingston contracting groups methods appears most useful. The empirical methods for discovering or adjusting standards are useful only to the extent that they call attention to the relative costs of the two types of errors.

3. Program evaluation

Standards impose an artifical dichotomy on data, and thus much information is lost about performance along the continuum in question. Shepard (1980, p. 468) states what many of us have believed and said for years: "Standards should not be used to interpret test data regarding the worth of educational programs."

V. Current Practices

Procedures used in Setting Standards*

Procedure	State	Local
Administrative Decision	5	6
Contrasting Groups	2	3
Nedelsky/Angoff	1	2
Field Test Results and/or Other Statistical Procedures	9	7
Competency Definition	3	2

*From National Evaluation Systems, 1979. The reader is referred to this report for additional information.

VI. Summary

This presentation has covered a lot of material. Rather than review it here, I shall simply present a list of "DO NOTS" and "DOs" for setting cutting scores.

DO NOTS

1. Do not set cutting scores before building the items.
2. Do not set cutting scores before gathering some empirical evidence on item difficulty from an appropriate sample of students.
3. Do not set cutting scores without some empirical evidence regarding the teachability of the material and the educational costs associated with the instruction.
4. Do not set cutting scores without explicit consideration by representatives of parents and educators of the relative costs of false positives and false negatives.
5. Do not set cutting scores which take effect the first year the test is administered.
6. Do not conclude that more rememdiation is needed in one basic skill than another based on the different proportions of "pass" scores on two non-equated tests.
7. Do not suggest to the public that evidence of minimum performance is sufficient (Porter (1978) published a news release regarding the proportion of students who received "acceptable" scores in Michigan".)
8. Do not assume that one can not or should not report scores in a more continuous fashion even if some arbitrary cut off point has been established.

DOs

1. Do consider using more than test information for making important decisions. If test scores are combined with other data (in a multiple-regression sense) consider using the obtained raw score (or continuous scaled score transformation) rather than the artifically dichotomized value.
2. Do remember that cutting scores can and probably do change over time.

I first presented the above list at the Twelfth National Symposium for Professionals in Evaluation and Research in Cincinnati on October 17, 1978. The fact that I still agree with it and that therefore I have made no observable growth troubles me not - we all know about the unreliability of gain scores!

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.

Berk, R.A. (ed) Criterion-Referenced Measurement: The state of the art. Baltimore: John Hopkins University Press, 1980.

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

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

Ebel, Robert L. 1978. The Case for Minimum Competency Testing. Phi Delta Kappan, 59, 8, 546-549.

Ebel, Robert L. Essentials fo educational measurement. Englewood Cliff, NJ: Prentice-Hall, 1972,

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

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

Gorth, W.P., and Perkins, M.R. A Study of Minimum Competency Testing Programs: Final Program Development Resource Document. Amherst, MA: National Evaluation Systems, December, 1979.

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

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

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

Journal of Educational Measurement, 1978, 15, 4, 237-319.

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

Judges 12: 5-6. The Living Bible.

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

Koffler, S.L. A comparison of approaches for setting proficiency standards. Journal of educational Measurement, 1980, 17, 167-168.

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

Mehrens, William A., "The Technology of Competency Measurement" in R.B. Lugle, M.R. Carroll, & W.J. Gephart (Eds.), Assessment of Student Competence Bloomington, IN: Phi Delta Kappa, 1979.

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

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

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

Pipho, Chris. 1978. "Minimum Competency Testing in 1978: A Look at State Standards." Phi Delta Kappan, 59, 9, 585-597.

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

Porter, John W. March 21, 1978. News Release.

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

Shepard, L. Standard setting issues and methods. Applied Psychological Measurement. 1980, 4, 4, 447-467.

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

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

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