Passing Scores

A Manual for Setting Standards of Performance on Educational and Occupational Tests

Samuel A. Livingston
Michael J. Zieky

Passing Scores

A Manual for
Setting Standards
of Performance
on Educational
and
Occupational Tests

Samuel A. Livingston
Michael J. Zieky

Authors' Note

We thank William H. Angoff, Ronald A. Berk, Carol A. Dwyer, Robert L. Ebel, John J. Fremer, Ronald K. Hambleton, Richard M. Jaeger, Robert L. Linn, John A. Meskauskas, W. James Popham, and Benjamin Shimberg for their many helpful comments on an earlier draft of this manual.

The opinions we have expressed in this manual are our own and do not necessarily reflect the opinions of our reviewers or the position of Educational Testing Service.

Table of Contents

Overview

This manual is written for the person who will be responsible for choosing the passing score on an educational or occupational test. Our purpose in writing the manual is to help you select and apply a method for choosing the passing score. Therefore, we have tried to concentrate on practical advice, rather than discussions of theory or descriptions of research findings. For the reader who is interested in those topics, we have included a brief bibliography at the end of the manual. The manual itself is divided into seven sections:

1. *Decisions, Standards, and Judgments:* some key concepts and some things to consider in choosing a method for choosing the passing score;
2. *Methods Based on Judgments About Test Questions:* a how-to-do-it section;
3. *Methods Based on Judgments About Individual Test-Takers:* another how-to-do-it section;
4. *Methods Based on Judgments About a Group of Test-Takers:* yet another how-to-do-it section;
5. *Choosing a Standard-Setting Method:* our recommendations for choosing among the methods presented in the previous three sections;
6. *Social and Political Issues:* a brief discussion of some sources of controversy over passing scores;
7. *Helpful Hints:* practical advice not included in the previous sections.

Decisions, Standards, and Judgments

Decisions

A test score is a piece of information about a person. How can you use that information to make a decision? One way is to consider each person's test score along with other information about that person, apply your own judgment, and make the decision. This case-by-case method of decision-making has some important advantages. Because you do not have to specify your criteria for the decision in advance, you can take account of any relevant information you may have about the test-taker, even if you did not originally plan to use it. Case-by-case decision-making offers each test-taker the chance to be considered individually as a whole person. However, it also has some serious drawbacks. It is subjective, in that two different decision-makers can arrive at different decisions on the basis of the same information. You cannot adequately describe your criteria for the decision in the form of a statement to the test-takers and other interested persons. In short, case-by case decision-making offers no assurance to the test-takers that they will be treated fairly. As a result, it can leave you open to charges of favoritism, prejudice, or arbitrary and capricious actions. For these reasons, you may prefer to use a *decision rule* that you will apply in the same way to all test-takers. Your decision rule will specify what information you will be using and how you will use it in making decisions about individual test-takers.

One very simple and very common type of decision rule is to classify the test-takers into two groups: a higher-scoring group and a lower-scoring group. Decision rules of this type are used in many different testing situations. Here are only a few examples:

The higher-scoring group will go on to another unit of instruction; the lower-scoring group will repeat the previous unit.

The higher-scoring group will receive a diploma or certificate; the lower-scoring group will not.

The higher-scoring group will be licensed to practice a profession; the lower-scoring group will not.

The lower-scoring group will receive some kind of special remedial instruction; the higher-scoring group will not.

The higher-scoring group will be admitted to a training program; the lower-scoring group will not.

The higher-scoring group will be given credit for a college course without taking the course; the lower-scoring group will not.

Even when more complicated decision rules are used, part of the rule often involves classifying test-takers into a higher-scoring group and a lower-scoring group. For example, a professional certifying board might decide to grant certification only to persons who have completed an accredited training program and have at least two years' experience in the profession and earn at least a specified score on a certification test.

To use a test score in these types of decision rules, you must choose the test score that will separate the higher-scoring group from the lower-scoring group. The purpose of this manual is to help you make that choice by describing several methods that you can use.

In this manual, we will use the traditional terms "pass" and "fail" to indicate the placing of a test-taker into the higher-scoring group or the lower-scoring group. We will refer to the score that separates the two groups as the "passing score." We realize that these terms will be inappropriate for some testing situations. However, we believe that our manual will be more useful if we use these concise and familiar terms.

Standards

A standard is an answer to the question, "How much is enough?" There are standards for many kinds of things, including the purity of food products, the effectiveness of fire extinguishers, and the cleanliness of auto exhaust fumes. When you choose a passing score, you are setting a standard for performance on a test.

Choosing the passing score would not be a problem if the test-takers' scores always fell neatly into two groups, one group of nearly perfect scores and one group of scores at or near the chance level. Unfortunately, in the real world of testing, we rarely get such clear-cut results. We have to face the difficult task of deciding how much is enough.

Standards can be either absolute or relative. A relative standard depends on comparisons between individuals; an absolute standard does not. In testing, a relative standard depends on comparisons among the test-takers. The question, "How good is good enough?" is answered in

terms of the test-takers' scores. For example, consider the following statements:

1. "If your score is in the top 5 percent of the group, it is good enough."
2. "If your score is above the average score of the group, it is good enough."
3. "If your score is not more than 20 points below the average score of the group, it is good enough."
4. "If your score is not in the bottom 2 percent of the group, it is good enough."

Each of these four statements expresses a relative standard. In each case, "good enough" is defined in terms of the scores of the test-takers. An individual test-taker's score will be compared against a standard that depends on the scores of the other test-takers. The higher the other test-takers' scores are, the higher the standard will be. In contrast, an absolute standard is one that does not depend on the performance of the test-takers who will be measured against it. For the person who takes a test that will be used with an absolute standard, it does not matter how well the other test-takers do, because their scores will not affect the standard.

To know whether a passing score represents an absolute standard or a relative standard, you need to know whether the test scores are expressed in absolute or relative terms. To say, "The passing score is 60 out of a possible 100," tells you little, unless you know what "60" means. If it means "60 percent of the questions answered correctly," the passing score represents an absolute standard. If it means "better than 60 percent of all the test-takers," or "two standard deviations* below the average score of the test-takers," the passing score represents a relative standard.

Choosing a passing score to represent a relative standard is not difficult; you choose the score that passes the desired number or percentage of the test-takers. For example, if the test is being used to select students for an advanced course that is limited to thirty students, the passing score will be the score that passes exactly thirty students. This manual will concentrate on methods for choosing a passing score that represents an absolute standard.

*The "standard deviation" is a measure of how widely the scores of a group of test-takers are spread out along the test score scale.

Judgments

Any standard—absolute or relative—is based on some type of judgment. A standard is an answer to the question, "How good is good enough?" and this question can be answered only by someone's judgment. The choice of a passing score will involve judgments at some point in the process. It is important that these judgments be

(1) made by persons who are qualified to make them;

(2) meaningful to the persons who are making them; and

(3) made in a way that takes into account the purpose of the test.

These three requirements are interrelated. Different methods for choosing a passing score require different types of judgments and, therefore, somewhat different qualifications for the judges. In describing each method, we will describe the necessary qualifications for the judges, and we will suggest ways to get them to keep the purpose of the test in mind when they are making their judgments.

Two Types of Wrong Decisions

Whenever you use a test to classify the test-takers into two groups, two kinds of wrong decisions can occur:

1. A test-taker who actually belongs in the lower group can get a score above the passing score;
2. A test-taker who actually belongs in the higher group can get a score below the passing score.

These wrong decisions occur because tests are almost never perfect measures of the knowledge and skills they are intended to measure. A test-taker's skills may vary from day to day and even from hour to hour. A test-taker may guess at some of the questions, and there is no way to distinguish a lucky guess from an answer that the test-taker really knew. For most tests, the questions or problems do not include every item of knowledge and every possible application of the skills that the test is intended to measure. The questions or problems are only a sample of all those that could have been included, and they may give a misleading picture of the skills of some of the test-takers.

For all these reasons, on most tests it is impossible to choose a passing score that will completely eliminate wrong decisions. You can reduce the chance of passing a test-taker who should fail, by using a higher passing score. However, by doing so you will increase the chance of failing a test-taker who should pass. Similarly, you can reduce the chance of failing a test-taker who should pass, by using a lower passing score, but you will increase the chance of passing a test-taker who should fail. Improving the test will reduce the number of wrong decisions but will not eliminate them entirely.

If either type of wrong decision were of no consequence, you would not need to use a test; you could simply pass everybody or fail everybody. For example, if passing an unqualified test-taker would do no harm at all, your best decision rule would be to pass everybody. The method you use to choose the passing score should take *both* types of possible wrong decisions into account.

Methods Based on Judgments About Test Questions

The three standard-setting methods we describe in this section of the manual are based on the concept of the "borderline" test-taker. This test-taker is the one whose knowledge and skills are on the borderline between the upper group and the lower group. These methods are based on the idea that, since the test-takers who belong in the upper group will tend to earn higher scores than those who belong in the lower group, the passing score should be the score that would be expected from a person whose skills are on the borderline.* The judgments these methods require are made in terms of the specific questions on the test.

These methods are relatively convenient and can be applied either before or after the test is administered. In addition, the process of making judgments about test questions focuses the judges' attention closely on the content of the test. Most important, the necessary data—judgments about test questions—can nearly always be obtained. However, the type of judgment these methods call for is not simply an evaluation of someone's performance that the judge can observe. Instead, these methods call for a much more difficult type of judgment. The judges must decide how a borderline test-taker would be likely to respond to each of the questions on the test. Because of the hypothetical nature of these judgments, we believe that these methods need a "reality check." If you use one of these methods, you should supplement it with some kind of information about the actual test performance of real test-takers, if you possibly can. And if this additional information clearly indicates that the results of the method do not describe the performance of a borderline test-taker, you should be prepared to admit that the method may not have worked well and to choose the passing score in some other way.

*The earliest article describing one of these methods (Nedelsky, 1954) referred to this person as the "F-D student."

Each of these methods consists of five basic steps:

1. Select the judges;
2. Define "borderline" knowledge and skills;
3. Train the judges in the use of the method you have chosen;
4. Collect judgments;
5. Combine the judgments to choose a passing score.

The first two steps are the same for all methods. The remaining steps differ.

The first step in any of these methods is to select the judges. The judges must be qualified to decide what level of the knowledge or skills measured by the test is necessary. For example, if a test of occupational knowledge is being used as a requirement for a nuclear power plant operator's license, the judges must be qualifed to decide how much knowledge is necessary to protect the public against operator errors that could result in a nuclear accident. If a reading test is being used as a requirement for high school graduation, the judges must be qualified to decide what a high school diploma should indicate about a person's reading ability. In some cases, only a few people may be qualified to serve as judges; in other cases, many may be qualified. If only a few people have the necessary qualifications, and if it is possible for all of them to participate as judges, try to include them all. Otherwise, try to make sure that the judges who participate are typical of all persons qualified to be judges. All important points of view should be represented on the panel of judges.

How many judges should you select? If you have too few, the process may be too greatly influenced by one or two individuals with unusually high or unusually low standards. In this respect, the more judges, the better. But the more judges you already have, the less you will gain from adding one more judge. We have used these methods with as few as five judges, but in these cases, the results were to be taken as a recommendation, not as a final determination. We suggest you try to get more if you possibly can.

Although it is possible to apply these methods without having the judges communicate directly with each other, we strongly recommend that you bring the judges together at a meeting. (If you have more than 20 judges, we suggest you divide them into smaller groups and work with each group separately.) At this meeting, you can have the judges define "borderline" knowledge and skills, and you can train the judges in applying whichever passing score selection method you have chosen. To define "borderline" knowledge and skills, first make sure the judges

understand what the test measures and how the test scores will be used. Then ask the judges to describe, in their own words, a person whose knowledge and skills would represent the borderline between acceptable and unacceptable levels of the knowledge and skills the test measures. The judges may find it convenient to describe the performance of specific people they have worked with, whom they would classify as "borderline." You can help the process along by asking appropriate questions. For example, if the test is a reading comprehension test that is being used to identify high school students who need further instruction in reading, you might ask, "Should the borderline test-taker be able to find specific information in a newspaper article? To distinguish statements of fact from statements of opinion? Should the borderline test-taker be able to recognize the main idea of a paragraph, stated in different words, if the paragraph is from a *Reader's Digest* article? How about a paragraph from an article in *Newsweek*? How about *Scientific American*?"

Allow the judges plenty of time to agree on a definition of borderline knowledge and skills. If there are strong differences of opinion that cannot be resolved by a compromise, you may have to proceed without a single definition that the entire panel of judges can agree on. But try to get agreement if you possibly can. When the judges have agreed on a definition, write it down, complete with examples, so you will have a statement in words of the standard that the passing score is supposed to represent.

From this point on, the methods differ. The three methods we will describe are named for the people who first suggested them in books and articles about educational measurement. The methods are known as "Nedelsky's method," "Angoff's method," and "Ebel's method." Each of the three methods requires a different type of judgment.

Nedelsky's Method

This method, suggested by Leo Nedelsky in 1954, can be used only with multiple-choice tests, since it requires a judgment about each possible wrong answer. The judge's task is to look at the question and identify the wrong answers that a borderline test-taker would be able to recognize as wrong, that is, as not the best of the answers presented. For example, consider the following question from a test of language skills. The test-taker's task is to choose the word or phrase that best completes the sentence.

"My music teacher thinks that Marian Anderson sings _____ any other contralto he has ever heard."

(A) more well than (B) better than

(C) the best of (D) more better over

A judge might decide that the borderline test-taker would be able to eliminate wrong answers A and D. But the judge might decide that the choice between wrong answer C and the correct answer B is too difficult for the borderline test-taker. The judge would then identify answers A and D as being so clearly wrong that the borderline test-taker would be able to recognize them as wrong.

Collecting the Judgments

Should the judges make their judgments individually or try to reach a consensus? The method seems to work fairly well either way, if the number of judges is not too large. But even with a small number of judges, it may take some time to get a consensus on each test question, and with more judges, it will be even harder to get them to agree. Yet, we believe that the judges can make more valid judgments if they share information and opinions with each other. Therefore, we recommend the following group procedure:

1. Have the judges make a set of preliminary judgments for all the questions, working individually and using a pencil to mark the wrong answers the borderline test-taker would be able to eliminate.
2. Conduct a *brief* discussion of each question, using the following format:
   a. Focus the judges' attention on the first wrong answer. Ask how many of them thought the borderline test-taker would be able to eliminate it as not the best answer, and how many did not think so.
   b. If the judges are not unanimous, ask one judge who marked the answer to explain why. Then ask one judge who did not mark that answer to explain why not. Do not try to reach agreement; just allow each point of view to be heard. The judges may or may not be swayed by the comments of their colleagues. Tell the judges they may change their judgments if they want to. Make sure the judges understand that their judgments are supposed to describe the performance of a *borderline* test-taker.
   c. Go on to the next wrong answer.
3. After all the questions have been discussed in this manner, ask the judges to review their decisions and make sure they have marked *all* the wrong answers they intended to mark and *only* those answers.

4. Collect the judgments.

To save time, you can use a shortcut version of this technique in which you consider each question as a whole:

1. Ask how many judges eliminated all the wrong answers.
2. Ask how many judges eliminated the first wrong answer, how many eliminated the second wrong answer, and so on.
3. Ask for one of the judges to explain his or her reasoning in deciding which wrong answers to eliminate.
4. Ask for one of the judges who made a different decision to explain his or her reasoning.
5. Allow discussion as long as the discussion seems to be productive. Then remind the judges that they can change their judgments if they want to.
6. Go on to the next question.

You may find it useful to begin by discussing each wrong answer and then switch, after a few questions, to discussing the question as a whole.

One limitation of this procedure is that it requires all the judges to make their judgments at the same time and place. Another limitation is that, even with the shortcut, it is fairly slow (though not nearly as slow as trying to get a group consensus on each question). For either of these reasons, you may find it necessary to have the judges make their judgments individually, without communicating with each other. If you do, remember that making this type of judgment will probably be an unfamiliar task for the judges. If possible, you should give them the chance to practice the judging task on a sample of the questions and discuss their work with each other before judging the rest of the questions. (This is the procedure Nedelsky recommended.)

Some types of multiple-choice questions present problems in using Nedelsky's method. One type that can cause problems is the negatively worded question, like the following example:

Which of the following foods is *not* a source of vitamin C?
(A) milk (B) orange juice (C) raw cabbage (D) baked potatoes

In deciding what wrong answers to mark, the judge must remember that the *better* a source of vitamin C a food is, the *worse* an answer to the question it is, and therefore the *more* likely the borderline test-taker would be to recognize it as wrong.

Another type of question that can cause problems with Nedelsky's

method is the "multiple true-false" question, such as the following example:

Which country or countries did the United States fight against during World War II?

I. Germany
II. Russia
III. Italy
IV. Japan

(A) I only (B) II only (C) I and IV only
(D) I, III, and IV only (E) I, II, III, and IV

This question is really four true-false questions, and the judge should deal with it that way. First the judge should decide which of the numbered choices the borderline test-taker would identify as correct, which choices the borderline test-taker would identify as incorrect, and which choices the borderline test-taker would be unsure about. Then the judge can figure out which of the answer choices (A, B, C, D, E) the borderline test-taker could eliminate. In the example, suppose the judge decides that the borderline test-taker would know that I (Germany) and IV (Japan) are correct and that II (Russia) is wrong. Then the borderline test-taker could eliminate answer choice A, because it does not include Japan, choice B, because it does not include Germany or Japan and does include Russia, and choice E, because it includes Russia.

If you decide to use Nedelsky's method with a test that contains negatively worded questions or multiple true-false questions such as those in the examples above, be sure to give the judges plenty of practice at judging those kinds of questions before they begin making their judgments individually. Make sure they can follow the logic of the judging process. When they have finished making their judgments individually, ask them to explain the reasons for their judgments on at least some of those questions, to make sure their marks are what they really intended.

Another type of question that can present difficulties in using Nedelsky's method is the question that requires the test-taker to do some mathematical computation. The wrong answer choices to these questions usually are the results of common mistakes. The difficulties arise because the type of mistake that a wrong answer choice indicates is not always obvious. Therefore, the judges may have a hard time deciding whether or not a borderline test-taker would have selected a particular wrong answer. Even the best qualified judges may find it time-consuming to figure out what kind of mistake would lead to each wrong answer. You can avoid this problem by giving the judges a copy of the test that shows the types of mistakes that lead to each wrong answer choice. For example, consider the following question:

A worker dropped a hammer off the roof of a building 36 feet high. How long did it take the hammer to reach the ground? (Use $g = 32$.)
(A) 1.06 seconds (B) 1.125 seconds
(C) 1.5 seconds (D) 2.25 seconds

A judge might know that the correct formula is $s = 1/2\ gt^2$, which leads to answer C, and yet have trouble figuring out where the wrong answers A, B, and D came from. The judge's task would be much easier and faster if the answers on his or her copy of the test were marked as follows:

(A) 1.06 seconds *($s = gt^2$)*
(B) 1.125 seconds *($s = gt$)*
(C) 1.5 seconds *($s = 1/2\ gt^2$)*
(D) 2.25 seconds *($s = 1/2\ gt$)*

One important issue in the application of Nedelsky's method (and Angoff's and Ebel's methods, also) is whether or not to tell the judges the correct answers to the test questions. Giving the judges the correct answers may make the questions seem easier than they are and, therefore, bias the judges in the direction of a higher cutoff score. If you do not give the judges the correct answers, they may judge some of the correct answers to be wrong answers that a borderline test-taker would eliminate, but this information can be valuable. If several judges eliminate the correct answer to the same question, that question may be defective. And if one judge eliminates many of the correct answers, that judge may be unqualified.

However, if you do not give the judges the correct answers, the judges may feel that they are being tested and may forget that their judgments are supposed to indicate the responses of a borderline test-taker. In addition, the judging process will surely take longer if the judges have to take the extra step of figuring out the right answer to each question. A good solution, if your situation permits it, is to have the judges take the test *before* the judging session and then give them the correct answers to use while they are actually making their judgments.

Choosing the Passing Score

Nedelsky's method is based on the idea that the borderline test-taker responds to a multiple-choice question by first eliminating the answers he or she recognizes as wrong and then guessing at random from the remaining answers. If the test is to be scored without a correction for

guessing, it is relatively easy to find the score that such a test-taker would be expected to get, by applying the following rules:

1. Under Nedelsky's method, the test-taker's expected score for any question is 1 divided by the number of answers the test-taker has to guess from.
2. To find a test-taker's expected score for the whole test, add up that test-taker's expected scores for all the individual questions.

For example, if the borderline test-taker has eliminated all but three possible answers, he or she has one chance in three of choosing the correct answer. Therefore, his or her expected score for that question is 1 divided by 3, or .33. Table 1 shows an example of these calculations for one judge's judgments on a ten-question test.

If the test will be scored with a correction for guessing, an additional calculation is necessary. This calculation is explained in the Appendix.

The calculations we have just described will give you a separate result for each individual judge. How should you combine these scores? One way is simply to average the scores in the usual way: add them up and

Table 1. Example of calculations for Nedelsky's method applied to a test scored without correction for guessing

Question	Answers*	Number of answers *not* eliminated	Expected score
1	A (B) ~~C~~ ~~D~~ ~~E~~	2	1/2 = .50
2	~~A~~ ~~B~~ ~~C~~ ~~D~~ (E)	1	1/1 = 1.00
3	~~A~~ ~~B~~ C (D) ~~E~~	2	1/2 = .50
4	A ~~B~~ C (D) ~~E~~	3	1/3 = .33
5	(A) ~~B~~ ~~C~~ ~~D~~ ~~E~~	1	1/1 = 1.00
6	A B (C) D E	5	1/5 = .20
7	A B C ~~D~~ (E)	4	1/4 = .25
8	(A) B ~~C~~ D E	4	1/4 = .25
9	A (B) C D E	5	1/5 = .20
10	A (B) C D E	5	1/5 = .20
			Sum = 4.43
Expected total score = 4.43			

*A circle indicates the correct answer; an X indicates an answer the borderline test-taker would eliminate.

divide by the number of judges. This type of average is called the *mean*. The disadvantage of using the mean is that it allows one judge with a very high or very low passing score to have a large influence on the result. A second way to combine the scores is to take the *median*. To find the median, first place the scores in order from highest to lowest. (If two judges arrive at the same score, be sure to list it twice, once for each judge.) If the number of judges is an odd number, the median is simply the middle score. If the number of judges is even, the median is halfway between the two middle scores. The disadvantage of using the median is that it disregards a great deal of information by focusing entirely on the middle score. A third way to combine the scores represents a compromise between the mean and the median. It is called the *trimmed mean*. To compute the trimmed mean, simply eliminate the highest and lowest scores and average the remaining scores in the usual way. Depending on the number of judges, you may choose to eliminate the highest two scores and the lowest two scores, or the highest and lowest three scores, or more. How much "trimming" to do is up to you.* If you are going to use the trimmed mean for averaging the scores, you should let the judges know this fact *before* you calculate the passing score from their judgments. Otherwise, the judges with the highest and lowest standards may suspect that you are discriminating against them. Table 2 shows an example of these three ways of combining the scores from the individual judges to choose a passing score. This example was constructed to show a case in which the three ways of combining scores produce very different results. In most cases the differences will not be as large as they are in Table 2.

Table 2. Example of three ways to combine scores from individual judges

Judge 1 (highest)	92.50	
Judge 2	77.25	Judge 2 77.25
Judge 3	67.00	Judge 3 67.00
Judge 4	66.67	Judge 4 66.67
Judge 5 (lowest)	65.33	
	Sum = 368.75	Sum = 210.92

Mean = 368.75 ÷ 5 = **73.75**
Median = 3rd highest = **67.00**
Trimmed Mean = 210.92 ÷ 3 = **70.31**

*One fairly common practice is to eliminate the highest 25 percent and the lowest 25 percent of the scores and average the middle 50 percent. The resulting statistic is called the "midmean."

When you have collected the judgments, computed the resulting score for each judge, and combined the results, you will have a consensus judgment of the score that a borderline test-taker would be expected to get on the test. Of course, even if this judgment is correct, not every borderline test-taker would get this exact score every time he or she takes the test. Rather, this expected score represents the score that is typical of a borderline test-taker's performance. If you choose this score as the passing score, a borderline test-taker should have a 50 percent chance of passing the test (if the Nedelsky-type judgments actually do describe the way such a test-taker would perform on the test). Therefore, in a fairly large group of borderline test-takers, about half would pass the test and about half would fail.

Angoff's Method

This method, suggested by William H. Angoff in 1971, is similar to Nedelsky's method, but it can be used with tests that are not multiple-choice. In Angoff's method, the passing score is computed from the expected scores for the individual questions, as in Nedelsky's method. However, Angoff's method does not require the judge to consider each possible wrong answer separately. Instead, the judge considers each question as a whole and makes a judgment of the probability that a borderline test-taker would answer the question correctly. This task may be difficult for some judges. If the judges are not comfortable about making judgments in terms of probabilities, ask them to imagine a group of 100 borderline test-takers and decide how many of them would answer the question correctly. Obviously, the easier the question, the higher this number will be. The probability must be between .00 and 1.00. If the questions are multiple-choice, the probability should ordinarily be at least as large as the chance of guessing the correct answer by blind luck (that is, 1.00 divided by the number of choices).

Collecting the Judgments

Should the judges make their judgments individually or try to reach a consensus? Again, we recommend a compromise procedure:

1. Have the judges make preliminary judgments for the first few questions only.
2. Conduct a brief discussion of each of these questions, using the following format:

a. Have each judge announce his or her choice of a probability for each question. Write these numbers on a blackboard or a large sheet of paper so all the judges can see them. If the numbers are all similar (e.g., within 10 or 15 percentage points), go on to the next question.

b. If the numbers are not all similar, ask for a judge who chose one of the highest numbers to explain the reasons for choosing a high probability. Then ask for a judge who chose one of the lowest numbers to explain the reasons for choosing a low probability.

c. Tell the judges they can change their judgments if they want to. Make sure the judges understand that their judgments are supposed to describe the performance of *borderline* test-takers.

3. After discussing the first few questions, have the judges make preliminary judgments for the remaining questions.

4. Discuss the remaining questions as in step 2, and give the judges a chance to change their judgments if they want to.

5. Collect the judgments.

Some people have used a modification of Angoff's method in which the judges are presented with a selection of probabilities in multiple-choice format and asked to circle one of the choices. We do not recommend this method, for two reasons. First, it can bias the judges' choices, particularly if the choices at one end of the scale are very limited. For example, suppose the judges are required to choose from the following list of probabilities:

.10 .20 .30 .40 .50 .75

A judge who thinks that all or nearly all borderline test-takers would answer the question correctly has no way to express that opinion. Second, limiting the judges' choice of probabilities is contrary to the logic of Angoff's method. If you believe that the judges can make valid probability judgments, you have no reason to restrict their choice. If you do not believe the judges can make valid probability judgments, you should not be using Angoff's method. The restricted choice makes sense only if you believe that the judges can make valid probability judgments with this kind of prompting but not without it.

Choosing the Passing Score

Finding the expected test score for a borderline test-taker is done in basically the same way as in Nedelsky's method. If the test is scored without a correction for guessing, the probability of a correct answer is the test-taker's expected score for that question. Simply add the probabili-

ties for the individual questions to get each judge's estimate of the borderline test-taker's expected score for the whole test. Table 3 shows an example. (If the test is scored with a correction for guessing, you must do the additional calculation shown in the Appendix.) You can combine the scores you have computed for the individual judges in the same way as for Nedelsky's method, by computing the mean, or the median, or the trimmed mean (see pages 22-23).

Table 3. Example of calculations for Angoff's method applied to a test scored without correction for guessing

Question	Probability of Correct Answer
1	.95
2	.80
3	.90
4	.60
5	.75
6	.40
7	.50
8	.25
9	.25
10	.40
	Sum = 5.80
Expected total score = 5.80	

Ebel's Method

Unlike the previous two methods, Ebel's method is a two-stage procedure. Each judge first classifies the questions into groups and then makes a single numerical judgment for each group of questions. The classification of questions into groups is based on two kinds of judgments about each question: a judgment of its *difficulty* and a judgment of its *relevance* (or importance). Ebel suggested three difficulty levels, labeled "easy," "medium," and "hard," and four relevance categories, labeled "essential," "important," "acceptable," and "questionable." The judge's first task is to classify all the questions in the test, which will result in a classification table similar to Table 4. (If you have statistics indicating

the difficulty of each question, you may want to make this information available to the judges to help them make the judgments of difficulty.)

The judge's second task is to make judgments about the performance of a borderline test-taker. The judge must make one such judgment for each of the 12 blocks of the classification table (except for those that are empty). That is, the judge must make one judgment for the questions classified "essential, easy," another for the questions classified "essential, medium," and so on, all the way down to "questionable, hard." The judgment consists of an answer to the question: "If a borderline test-taker had to answer a large number of questions like these, what percentage would he or she answer correctly?" Table 4 includes examples of these judgments.

Table 4. Example of classification of questions (stage 1) and judgments (stage 2) in Ebel's method

	Difficulty:		
Relevance:	**Easy**	**Medium**	**Hard**
Essential	Questions #1,4,7,8,13 Judgment: 95% correct	Questions #11,15,22 Judgment: 85% correct	Question #21 Judgment: 80% correct
Important	Questions #2,6,9 Judgment: 90% correct	Questions #10,14,20 Judgment: 75% correct	Questions #16,25 Judgment: 60% correct
Acceptable	Question #5 Judgment: 80% correct	Questions #12,18 Judgment: 55% correct	Questions #19,23 Judgment: 35% correct
Questionable	Question #3 Judgment: 50% correct	Questions: none No judgment needed	Questions #17,24 Judgment: 20% correct

Collecting the Judgments

The group procedure that we recommend for Nedelsky's method and Angoff's method can be adapted for Ebel's method. However, it will be more complicated, because the judges must make two decisions about each test question—its difficulty and its relevance—and must then make a judgment about the borderline test-taker's performance on each of the 12 groups of questions. If you use this procedure for Ebel's method, we recommend applying it separately to each of the two stages of Ebel's method. The resulting procedure would be as follows:

1. Have the judges make a preliminary classification of the test questions into the 12 categories, working individually.

2. Conduct a brief discussion of each question, using the following format:
   a. Ask how many judges classified the question as "easy," as "medium," and as "hard." If the judges were not unanimous, ask one judge who classified the question as "easy" to explain why. Do the same for "medium" and "hard."
   b. Ask how many judges classified the question as "essential," as "important," as "acceptable," and as "questionable." If the judges are not unanimous, ask one judge who chose each category to explain why.
   c. Give the judges a chance to reclassify the question if they want to.
3. Have the judges make a preliminary judgment, for each of the 12 categories, of the percentage of such questions a borderline test-taker would answer correctly.
4. Conduct a brief discussion for each of the 12 categories, using the following format:
   a. Have each judge announce his or her choice of a percentage for that category.
   b. Ask a judge who chose one of the highest numbers to explain the reasons for choosing a high percentage. Then ask a judge who chose one of the lowest numbers to explain the reasons for choosing a low percentage.
   c. Tell the judges they may change their judgments if they want to. Make sure the judges understand that the judgments are supposed to describe the performance of a *borderline* test-taker.
5. Collect the judgments.

Choosing the Passing Score

To find the expected test score for a borderline test-taker, use the following procedure:

1. Multiply the judged percentage correct for the first category ("essential, easy") by the number of questions in that category to get the test-taker's expected score for the first category.
2. Repeat step 1 for each of the other 11 categories.
3. Add the expected scores for the twelve categories to get the expected score for the whole test.

Table 5 shows the calculations based on the classifications and judgments in Table 4. (If the test is scored with a correction for guessing, you must perform the additional calculation shown in the Appendix.) You

can combine the scores you have computed for the individual judges in the same way as for Nedelsky's method or Angoff's method, by computing the mean, or the median, or the trimmed mean (see pages 22-23).

Table 5. Example of calculations for Ebel's method applied to a test scored without correction for guessing

Category	Percentage Correct	Number of Questions	Expected score for category
Essential			
Easy	95	5	.95 × 5 = 4.75
Medium	85	3	.85 × 3 = 2.55
Hard	80	1	.80 × 1 = .80
Important			
Easy	90	3	.90 × 3 = 2.70
Medium	75	3	.75 × 3 = 2.25
Hard	60	2	.60 × 2 = 1.20
Acceptable			
Easy	80	1	.80 × 1 = .80
Medium	55	2	.55 × 2 = 1.10
Hard	35	2	.35 × 2 = .70
Questionable			
Easy	50	1	.50 × 1 = .50
Medium	*	0	.00
Hard	20	2	.20 × 2 = .40
			Sum = 17.75
Expected total score = 17.75			

*Information not needed—no questions classified into this category.

Methods Based on Judgments About Individual Test-Takers

The methods presented in this section are based on information about individual test-takers. They require two types of information about each test-taker: (1) the person's test score, and (2) a judgment of the adequacy of the test-taker's knowledge and skills. These methods include the "borderline-group" method, the "contrasting-groups" method, and a variation of the contrasting-groups method called the "up-and-down" method. The main advantage of these methods is that people in our society are accustomed to judging other people's skills as adequate or inadequate for some purpose—especially in educational and occupational settings. Teachers judge the skills of their students, supervisors judge the skills of the workers they supervise, and professionals judge the skills of their colleagues. Therefore, making this type of judgment is likely to be a familiar and meaningful task.

The judgments used in these methods should meet the following four requirements:

1. The judgments must be made by persons who are qualified to make them;
2. The judgments must be judgments of the knowledge and skills the test is intended to measure;
3. The judgments must reflect the test-takers' skills at the time of testing;
4. The judgments must reflect the judges' true opinions.

The first requirement applies to any method of choosing a passing score: the judgments must be made by qualified persons. With methods based on judgments of individual test-takers, two kinds of qualifications are necessary: (1) the judges must be able to determine each test-taker's knowledge and skills, and (2) the judges must know what level of knowledge and skill a person passing the test should have. It is important that the judges have both these qualifications. If you cannot find judges who have both, you may be able to design the standard-setting process so as to provide the information that the judges lack. That is, you can choose judges who are familiar with the test-takers' knowledge

and skills and make them aware of the level of knowledge and skills that will be required. Alternatively, you can choose judges who understand the level of knowledge and skills required and give them the opportunity to observe the test-takers' knowledge and skills.

If the test-takers are students, their teachers or instructors may be able to provide informed judgments of their knowledge or skills. In this case, it is a good idea to tell the teachers not to make any judgment of a student whose skills they have not had the chance to observe adequately. The same principle applies when you are asking supervisors to judge the workers they supervise, or when you are asking test-takers to judge their peers.

In some cases the test-takers themselves may provide the judgments of their own knowledge and skills. For example, suppose an instructor wants to use a math test to determine whether students' math skills are adequate for a technical training course. The instructor could give the test to all the students at the beginning of the course the first time it is given. After the students have progressed far enough in the course to need those skills, the instructor could ask the students to make a judgment: "Do you feel that your math skills at the time you began this course were adequate for the course?" The instructor could then use those judgments to set a passing score on the test for the next group of students applying for the course. Notice that in this example the students would meet both qualifications for judges: They would be aware of their own skills and of the level of skill required.

If the judges are not already familiar with the test-takers' knowledge and skills, you will have to give them a chance to observe a demonstration or an example of the product of each test-taker's knowledge and skills. For example, if the test-takers are x-ray technologists, the judges can observe their procedure and inspect some of the x-ray pictures they have taken. While you may not be able to arrange for observations of all the test-takers, you may be able to get observations of a sample of the test-takers.

What if the test itself is the best available indication of the test-takers' skills? In this case, the judges can base their judgments on an observation of the test-takers' actual test performance—not the test score, but the performance itself. For example, when an essay test is used to test students' writing skills, the judges can read the students' essays. For a test of foreign-language speaking ability or musical performance, the judges can listen to the actual performance, or a portion of it (either live or recorded). The same principle applies to any performance test that is objectively scored.

A second requirement is that the judgments must be based on the

skills and knowledge the test is intended to measure. The problem is that judgments of individuals' skills may be affected by factors that are irrelevant to the purpose of the test. For example, teachers who are asked to judge their students' skills in English composition may allow their judgments to be influenced by the students' understanding of literature, their penmanship, their punctuality in completing assignments, their class participation, and so on. Instructions to the judges can help to reduce the influence of these irrelevant factors. The judges must understand clearly which characteristics of the test-takers they should judge and which they should disregard.

A third requirement is that the judgments must reflect the test-takers' skills at the time of testing. If the judgments are based on the judges' familiarity with the test-takers' knowledge and skills, the judgments should be made as close to the time of testing as possible. If the judgments are based on a special observation, the performance that the judges observe should be done as close to the time of testing as possible. (If this performance is recorded in some way, it can be observed and judged at a later time.)

There is one exception to this requirement. If the test is intended to predict the test-takers' skills at some future time, then the judgments should be made at that future time. For example, if a test is intended to predict success in a training course, the judgments would have to be made at the end of the training course.

A fourth requirement is that the judgments must reflect the judges' true opinions. It is important to make sure that the judges have no personal incentive to be especially strict or especially lenient in judging the test-takers' skills. For example, when teachers are being asked to judge their students' skills, the teachers may suspect that their judgments will be used to evaluate the effectiveness of their teaching. The best precaution against this sort of misunderstanding is to make sure the judges understand how their judgments will be used. They should realize that by participating in the standard-setting exercise, they are assuring that the passing score will reflect their own individual standards.

We strongly recommend that the judges *not* know the test-takers' test scores until after the judging process is complete. Even if the judgments are based on a performance that is part of the test itself, they should be judgments of the performance, not of the test scores. The danger is that a judge who knows the test-takers' scores may use the scores of the first few test-takers to establish a standard and then judge the rest of the test-takers by comparing their test scores with those of the first few. If the first few test-takers are not typical, all of the remaining judgments will be distorted. But if the judges do not have access to the test scores, they will

have to judge each test-taker individually, and the standard-setting procedure will work the way it is supposed to.

The Borderline-Group Method

This method is based on the idea that the passing score should be the score that would be expected from a test-taker whose skills are "on the borderline"—not quite adequate and yet not really inadequate. In this respect it resembles the methods based on judgments of test questions. However, instead of asking the judges to make educated guesses about the way a borderline test-taker would perform, this method calls for the judges to identify actual test-takers as "borderline" in the knowledge and skills the test measures. The judges do not have to judge all of the test-takers or even a representative sample of them. They need only identify the ones who, in their judgment, best fit the definition of a borderline test-taker. You then set the passing score at the median score (the 50th percentile) of this "borderline group." The main advantage of this method is its simplicity. It is easy to use and easy to explain. The main disadvantage of this method is that borderline test-takers usually are a small percentage of all the test-takers. The judges may have trouble identifying test-takers who are truly "borderline."

You can apply the borderline-group method by the following sequence of steps:

1. Select the judges.
2. Define adequate, inadequate, and "borderline" levels of the skills and knowledge tested.
3. Identify "borderline" test-takers.
4. Obtain the test scores of the "borderline" test-takers.
5. Set the cutoff score at the median test score of the borderline group. This is the score that divides the group exactly in half, i.e., half the members above and half below.

The reason for using the median, rather than the mean (the usual "average"), is that the median is much less affected by a few extremely high or extremely low scores. This feature of the median is especially important for the borderline-group method, because a test-taker with a very high or very low score is likely to be someone who did not really belong in the borderline group.

If most of the test scores of the borderline group are clustered close together, then the method is working well. But if the scores of the borderline group are spread widely over the range of possible scores, then

the method is not working well. What can cause the borderline-group method to work poorly?

1. The borderline group may include many test-takers who do not belong in it. The judges may have identified several test-takers as "borderline" because their skills were difficult to judge.
2. The judges may be basing their judgments on something other than what the test measures.
3. The judges may differ considerably in their individual standards for judging the test-takers.

You may be able to avoid the first problem by reminding the judges not to include in the borderline group any test-takers whose skills they are not familiar with. You can minimize the second and third problems by giving the judges appropriate instructions and by getting them to agree with each other, before making their judgments, on a definition of "borderline" knowledge and skills.

The Contrasting-Groups Method

This method is based on the idea that the test-takers can be divided into two contrasting groups—a "qualified" group and an "unqualified" group—on the basis of the judgments of their knowledge and skills. Once you have divided the test-takers into these two groups, you can consider all the test-takers with a particular test score and ask, "Are the majority of them qualified or unqualified?" Most of the test-takers with very high scores will be in the "qualified" group. As you go down the score scale, the proportion of the test-takers who are "qualified" will decrease. At the lowest score levels, the "unqualified" test-takers will outnumber the "qualified" test-takers. One obvious choice for a passing score would be the score at which there are just as many "qualified" test-takers as "unqualified" test-takers.

In many cases it will not be practical to get judgments of all test-takers in the population. You may have to settle for judgments of a sample of the test-takers. How should you choose the sample? If you have to choose the sample of test-takers before you have given the test, you can choose them at random (for example, by lottery) from among all the people who will be taking the test. But if you can choose them after they have taken the test, there is a better way. You can choose the test-takers so that their scores are spread evenly throughout the portion of the score range where the passing score might possibly be located. For ex-

ample, on a 100-question test, you might choose 10 test-takers from each five-point score interval (31-35, 36-40, etc.). The important principle to remember is that the sample of test-takers you select *at each score level* must be representative of all the test-takers at their score level.

You can apply the contrasting-groups method by the following sequence of steps:

1. Select the judges.
2. Define adequate and inadequate levels of the knowledge and skills tested.
3. Select the sample of test-takers whose skills will be judged. (Omit this step if you can get judgments of all the test-takers.)
4. Obtain the test scores and the judgments of the test-takers you have selected. Do *not* let the judges know the test-takers' scores.
5. Divide the test-takers at each score level into "qualified" and "unqualified" groups on the basis of the judgments. Compute the percentage of the test-takers at each score level who are in the "qualified" group. (If you do not have several test-takers at each score level, combine score levels into larger intervals before you do this calculation.)
6. Use a "smoothing" method (explained below) to adjust the percentages you have computed.
7. Choose the passing score on the basis of the "smoothed" percentage.

"Smoothing" the Data

When you compute the percentage of the test-takers at each score level who are "qualified" (step 5 above), you may find that the percentage does not increase steadily from one level to the next. Instead, it may follow a zigzag pattern. For example, in Table 6, as you go down the test score scale, the percent qualified drops from 100 to 75, jumps to 95, drops to 60, rises to 69, drops steadily to 18, then jumps to 43, and so on. This kind of result is especially likely if the number of test-takers at each score level is small. It seems reasonable to assume that if you could get judgments of all possible test-takers, the percent-qualified would increase steadily from one score level to the next (possibly leveling off at the highest and lowest levels). What you need, then, is a way to adjust the percentages to bring them closer to what you would have found if you had obtained test scores and judgments of all possible test-takers.

The general term for adjustments of this kind is "smoothing." Figure 1

shows why. The solid line on the graph connects the actual observed percentages. The broken line connects the "smoothed" percentages. The broken line is "smoother" and, presumably, closer to the percentages that would be observed if a much larger group of test-takers had been judged.

Table 6. Data for examples of smoothing*

Test Score	Number of Test-Takers			Percent Qualified
	Qualified	Unqualified	Total	
96-100	5	0	5	100
91-95	3	1	4	75
86-90	6	2	8	75
81-85	18	1	19	95
76-80	17	3	20	85
71-75	15	10	25	60
66-70	20	9	29	69
61-65	7	8	15	47
56-60	6	17	23	26
51-55	2	9	11	18
46-50	6	8	14	43
41-45	2	4	6	33
36-40	2	12	14	14
31-36	0	7	7	0
0-30	0	3	3	0

*From W. Kastrinos and S. A. Livingston, *The Development of a Proficiency Examination for Dental Auxiliaries.* (Princeton, N.J.: Educational Testing Service, 1979), p. 64.

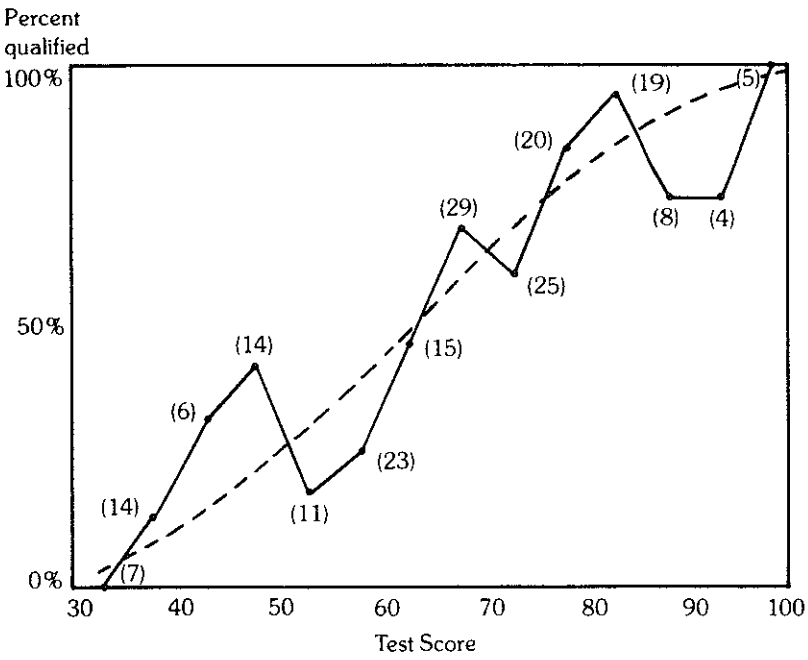

Figure 1. Example of of graphic smoothing.
(Numbers in parentheses indicate the number of test-takers at each test score level.)

There are several techniques for smoothing observed percentages. Some smoothing techniques involve complex calculations, but others are extremely simple. All smoothing methods are based on the idea that the judgments of test-takers at each test score level tell you something about the knowledge and skills of test-takers at nearby test score levels. One smoothing method that is easy to apply is to draw a graph like Figure 1, showing the percentages as points. Then try to draw a smooth curve that comes as close to the points as possible. If the number of test-takers varies from one level to the next, try to get the curve closer to the points that represent larger numbers of test-takers. This technique is called "graphic smoothing." It is somewhat subjective; that is, different people applying the method could come up with slightly different results. Nevertheless, it works well; that is, it produces results that are very similar to the results of the more objective methods of smoothing.

Another simple smoothing method is to replace the observed percentage at each test-score level with the average of the percentages for that score level and the two adjacent score levels. For example, in Table 6, the "smoothed" percent-qualified for test-score level 86-90 would be the average of the percentages for test-score levels 81-85, 86-90, and 91-95. This number would be the average of 95, 75, and 75, which is approximately 82. We would expect that in a very large group of test-takers with scores between 86 and 90, the percent judged to be qualified would be closer to 82 than to 75.

An improvement on this method is to weight each percentage by the number of test-takers at each score level. This procedure has the effect of combining the test-takers at the three score levels and computing the percent-qualified for this enlarged group. Table 7 illustrates this "moving

Table 7. Smoothing by "moving average"

Test Score	Number of test-takers Qualified	Total	"Smoothed" Percent Qualified
96-100	5	5	*
91-95	3	4	$\frac{5+3+6}{5+4+8} = 82\%$
86-100	6	8	$\frac{3+6+18}{4+8+19} = 87\%$
81-85	18	19	$\frac{6+18+17}{8+19+20} = 87\%$
76-80	17	20	$\frac{18+17+15}{19+20+25} = 78\%$
71-75	15	25	$\frac{17+15+20}{20+25+29} = 70\%$
66-60	20	29	$\frac{15+20+7}{25+29+15} = 61\%$
61-65	7	15	. . . and so on.

*This method cannot be used to estimate the percent-qualified at the lowest and highest test score levels.

average" method. The "moving average" cannot be computed at the very lowest and highest test-score levels, but this limitation should not often present a serious problem in setting cutoff scores. Notice that the results of this method are "smoother" than the original observed percentages shown in Table 1; that is, the percent-qualified does not change so abruptly from level to level. However, the smoothing did not remove all the inconsistencies; the smoothed percentage for test-score level 91-95 is still less than for the two score levels immediately below it.

Different smoothing methods can result in different passing scores. Although these differences will tend to be small, you may want to keep the process as objective as possible by specifying which smoothing method you will use before you collect the data. You may find that the resulting curve is not as smooth as you would like, but you will be protected against the charge that you deliberately chose a smoothing method that would produce a particular passing score.

Choosing the Passing Score

The final step in applying the contrasting-groups method is the choice of the passing score. One logical choice is the test score for which the "smoothed" percent-qualified is exactly 50 percent. At any lower test-score level, a test-taker is more likely to be judged unqualified than qualified, while the reverse is true at any higher test-score level. For the smoothed percentages indicated by the curve in Figure 1, this reasoning would lead to a passing score of approximately 65.

The rationale for setting the passing score at the test score that corresponds to a 50 percent chance of being judged as qualified is based on the assumption that the two types of possible wrong decisions about a test-taker are equally serious. But what if they are not? For example, what if it is twice as bad to pass an unqualified test-taker as it is to fail a qualified test-taker? In this case, the passing score should be higher, but how much higher? Statistical decision theory (which, at its simplest levels, is really common sense expressed in mathematical language) provides an answer to this question. The answer is based on the idea that your choice of a passing score should depend on the total harm from all the wrong decisions you can expect to make.

If it is twice as serious to pass an unqualified test-taker as it is to fail a qualified test-taker, then passing an unqualified test-taker would be exactly as bad as failing two qualified test-takers. The best choice for the passing score would be the test score at which there are exactly two qualified test-takers for every unqualified test-taker. This would be the test score that corresponds to 67 percent-qualified. By similar reason-

ing, if it were three times as bad to pass an unqualified test-taker as to fail a qualified test-taker, the passing score would be the test score at which qualified test-takers outnumber unqualified test-takers by three to one. That is, the passing score would be the test score that corresponds to 75 percent-qualified. On the other hand, failing a qualified test-taker might be the more serious of the two types of errors (for example, if you were testing to determine whether a student will receive an expensive remedial training program). In this case, you might want to lower the passing score to the test-score level where unqualified test-takers outnumber qualified test-takers by two to one or three to one.

In practice, you may find it simpler to ask yourself (and any other persons who are responsible for choosing the passing score) such questions as:

> "Suppose you had a group of 100 people and you knew that 50 were qualified and 50 were unqualified. If you had to pass all 100 or fail all 100, which would you do?"

If your answer would be "Fail them," then ask the same question for a group of 70 qualified persons and 30 unqualified persons. If your answer would now be "Pass them," ask the same question for a group of 60 qualified persons and 40 unqualified persons. Keep adjusting the percent-qualified in this way until you have found the value at which you cannot decide whether to pass the group or fail the group. The test score that corresponds to this percent-qualified will be the score at which you cannot decide whether a test-taker should pass or fail—that is, the passing score.

How Many Test-Takers?

One question that test users often ask about the contrasting-groups method is, "How many test-takers do I need?" The only honest answer to this question is, "It depends." Deciding how many test-takers to include in a contrasting-groups study generally involves a tradeoff between costs and benefits. The costs are those of getting the judgments. Judging more test-takers will require more time from the judges, or it may require you to select and train more judges. It may also require time from more of the test-takers. The benefits of a larger sample are better representation of the test-taker population and greater precision in determining the passing score. The degree of precision you can get with a given number of test-takers depends on several factors:

—the extent to which the test scores and the judgments both reflect the same abilities of the test-takers;

—the extent to which the test scores and the judgments are free of other influences;

—the consistency of the test-takers' performance:

—if different judges judge different test-takers, the extent to which the judges have the same standards;

—the consistency with which the judges apply their standards in judging the test-takers.

The degree of precision you need will depend on the number of people who will be affected by the choice of the passing score and on the consequences of passing or failing the test. It will also depend on how fine a distinction you are trying to make. A choice between passing scores of 3 and 4 on a five-point test is much easier to make than a choice between passing scores of 73 and 74 on a 100-point test.

One of us has used the contrasting-groups method with as few as 20 test-takers, but the circumstances of that study were somewhat unusual. Only seven test score levels were being considered as possibilities. Each test-taker was judged by eight judges, and the judgments were based on a sample of performance from the test itself. (It was a test of English-speaking proficiency for persons whose native language was not English.)* Most circumstances would call for judgments of a considerably larger number of test-takers.

The costs of getting judgments of individual test-takers, the precision that a given number of test-takers will provide, and the need for precision in setting the passing score will all vary from one testing situation to another. Therefore, we cannot prescribe a minimum number of test-takers that will apply to all testing situations. We can only suggest that you (1) include as many test-takers as you can afford to, and (2) consult a statistician for advice that will apply to your testing situation.

*For a description of this study, see Samuel A. Livingston, "Setting Standards of Speaking Proficiency," pp. 255-270 in *Direct Testing of Speaking Proficiency: Theory and Application*, J. L. D. Clark, editor. (Princeton, N.J.: Educational Testing Service, 1978).

The Up-and-Down Method

One problem that often makes it difficult to use the contrasting-groups method is the effort and expense involved in getting judgments of individual test-takers' skills. In many cases, the effort and expense depend directly on the number of individual test-takers to be judged. The more judgments, the greater the cost. Therefore, you will want to concentrate these valuable judgments in the part of the test-score range where you most need them—the part where about half the test-takers are qualified and half are not. But until you have collected the judgments, you will not know where this part of the score range is. Is there any way out of this dilemma? In some situations, the answer is "yes." If the test-takers take the test before the judgments of their skills are made, and if you can select the test-takers for judgment one at a time, you can use a variation of the contrasting-groups method called the "up-and-down method." The up-and-down method should work especially well where every test-taker's performance has been recorded and is available for judging, as in the case of a writing sample or an essay test. Here is how it works:

1. Select a test-taker with a test score near where you think the proper passing score might be. Get a judgment of this test-taker's skills.

2. If the first test-taker was judged to be qualified, choose next a test-taker with a somewhat lower test score. If the first test-taker was judged to be unqualified, choose next a test-taker with a somewhat higher test score. Get a judgment of the second test-taker's skills.

3. Repeat step 2, choosing the third test-taker on the basis of the judgment of the second test-taker. Continue by choosing each test-taker on the basis of the judgment of the previous test-taker.

Figure 2 illustrates an application of the up-and-down method. The letters Q and U in the figure represent judgments of the test-takers as being qualified or unqualified. Notice the way in which the method automatically tends to move down from test-score levels where all the test-takers are qualified and up from test-score levels where all the test-takers are unqualified. The scores of the test-takers selected will tend to concentrate in the range where a test-taker is about as likely to be qualified as to be unqualified—which is where the passing score should be.

To choose the passing score on the basis of data collected by the up-and-down method, you can simply take the average test score of the persons selected for judging, beginning just before the scores start to zig-zag and ending with the score of the next person who would have been judged if the procedure had continued. That is, disregard the first run of

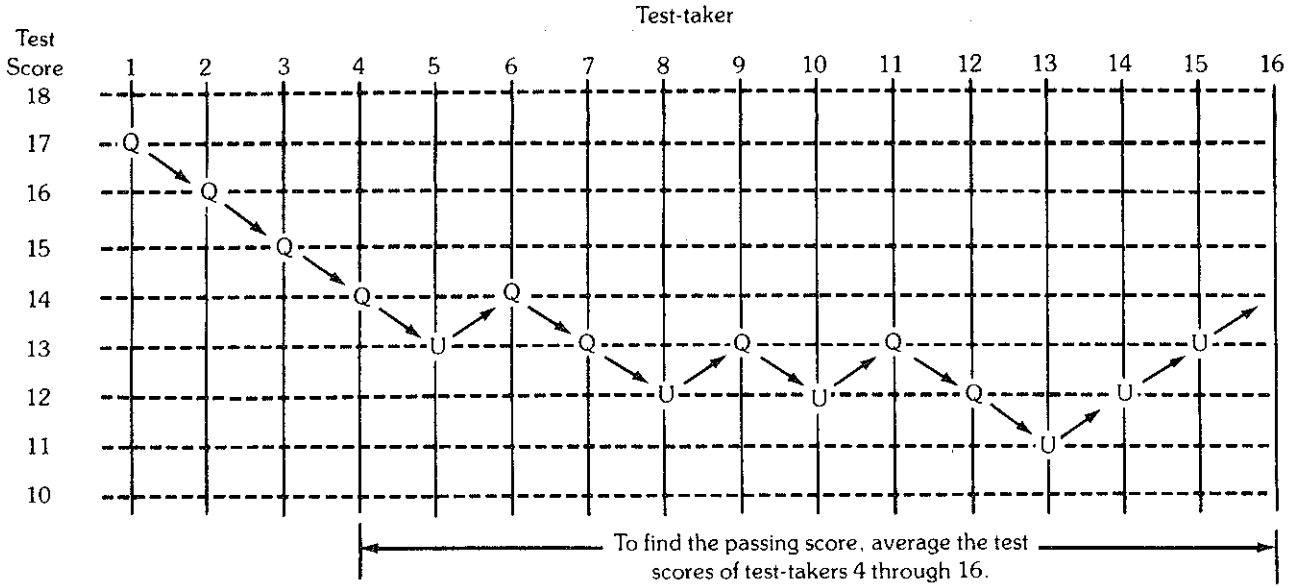

Figure 2. Example of the up-and-down method (hypothetical data)

qualified persons or of unqualified persons, except for the last person in that run. For example, in Figure 2, the first four test-takers were all judged to be qualified, so we would start with the fourth test-taker. The 16th test-taker was not actually judged, but we know that person's test score, so we include it in the average. The passing score would be the average score of test-takers 4 through 16, which is 12.8. Of course, in most situations you would want to get judgments of more than 15 test-takers.

A variation on the up-and-down method is to select more than one test-taker at a time. For example, you might select three test-takers at a time, all with test scores at the same level. If at least two of them are judged to be qualified, you would move down to a lower test-score level for the next three; otherwise you would move up to a higher level.

You can use this variation of the up-and-down method to find the test score for which the percent-qualified is something other than 50 percent. For example, suppose you want to find the score level at which two-thirds of the test-takers are qualified. You could select five test-takers at a time. If four or five (that is, more than two-thirds of the five) are judged qualified, you would move down to a lower test-score level for the next group of five; otherwise you would move up. A word of caution: If you are looking for some percentage other than 50 percent, you should *not* set the passing score by averaging the test scores of the persons you select. Instead, you should treat the data as you would in the regular contrasting-groups method: (1) compute the percent-qualified at each score level, (2) smooth the percentages if necessary, and (3) find the test-score level that corresponds to the percent-qualified you have chosen.

If you are using the up-and-down method to choose a passing score, it is important not to stop until you have observed several "reversals." A reversal is a change in direction, from up to down or vice versa. For example, in Figure 2, the reversals come after test-takers 5, 6, 8, 9, 10, 11, and 13. The importance of these reversals is that they will tend to come frequently in the range where the passing score should be. In other parts of the test-score range, there will be fewer reversals. The more reversals you have observed, the more likely it is that you have found the right portion of the test-score range.

How large should the steps be? That is, how far down the test-score scale should you move after a success, and how far up after a failure? The larger the steps, the more quickly you can find the part of the test-score range where the passing score should be. On the other hand, smaller steps will give you a more precise estimate once you reach that range. Therefore, we suggest the following procedure. Use large steps until you have observed at least five reversals. Then take one last large

step and switch to smaller steps. A large step might be one-eighth or one-tenth of the range of actual scores on the test (that is, of the difference between the highest and lowest of the test-takers' scores). A small step might be about half that size. For example, if the test-takers' scores range from 20 to 80, you might start with steps of 6 test-score points and then shift to steps of 3 test-score points, as in Figure 3.

One possible problem with the up-and-down method is that if the judges know you are using it, each judgment may be affected by the previous one. That is, if a judge knows that the test-taker now being judged had a higher test score than the previous one, the judge may be more inclined to judge the test-taker as qualified. We suggest that you not tell the judges what rule you are using to select the test-takers until the judging is finished. The judges may figure out the principle by themselves, but unless you tell them, they will not be sure you are following it consistently. Therefore, they will be more likely to continue to judge each test-taker as an individual.

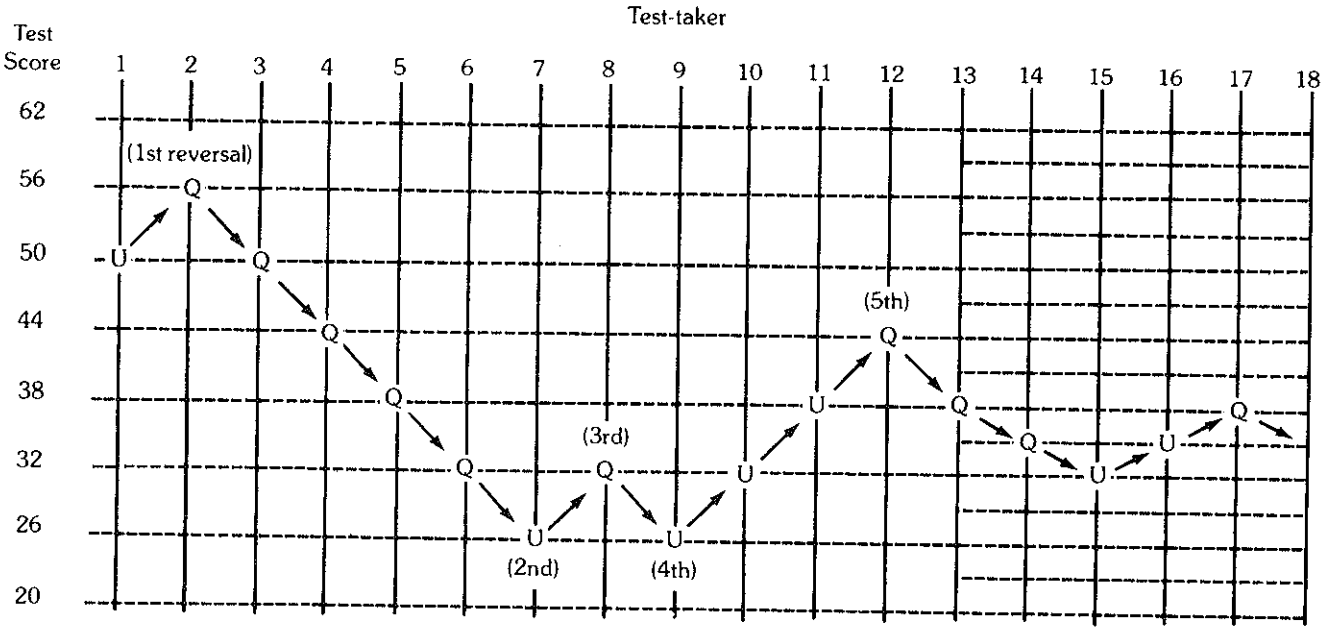

Figure 3. Example of the up-and-down method with a change in the step size (hypothetical data)

Methods Based on Judgments About a Group of Test-Takers

The methods described in this section are based on judgments about a group of test-takers—preferably a large group. This group is often called the *reference group*. The simplest of these methods, and the one with the most obvious justification, is to choose the passing score that would have passed a specified number (or a specified percentage) of the test-takers in the reference group. For example, if you have reason to believe that 85 percent of last year's test-takers were qualified, you can find the score that would have passed 85 percent of last year's test-takers and use that score as a passing score for this year's test-takers. (If the test changes from year to year, you will have to find the score on this year's test that would have passed 85 percent of last year's test-takers, by using a statistical technique called "equating."*) The judgment of the percentage of the test-takers in the reference group who were qualified leads directly to the choice of a passing score. This judgment should be based on some type of information *other* than the test scores.

Does a passing score chosen by this method represent an absolute standard or a relative standard? The answer to this question depends on the reference group. If the reference group is the group of test-takers the passing score will be applied to, then the standard is a relative standard. In this case a test-taker's relative standing in the group determines whether or not he or she passes the test. But if the reference group is a previous group of test-takers, it has the effect of setting an absolute standard. From the test-taker's point of view, the passing score has already been determined. Any test-taker who scores higher than that score will pass the test, even if the other test-takers all score higher still. And any

*For information on equating, see the chapter by W. H. Angoff cited in the bibliography of this manual.

test-taker who scores lower than the passing score will fail, no matter how poorly all the others do.*

You can apply this method by the following sequence of steps:

1. Identify the reference group.
2. Select the judges.
3. Define adequate and inadequate levels of the knowledge and skills tested.
4. Collect judgments of the percentage of the people in the reference group who have an adequate level of the knowledge and skills tested.
5. Choose the passing score.

Steps 1 and 2 are interdependent; your choice of a reference group will depend on your being able to find judges who can make a valid judgment about that group.

The reference group should be fairly large, so that the judgments of the percentage of the test-takers who are qualified will not depend heavily on one or two of the test-takers. You do not need to know the test scores of individual test-takers, but you do need to know how many test-takers in the group received each test score.

The judges must be able to judge how many (or what percentage) of the test-takers in the reference group are qualified in the knowledge and skills the test measures. Therefore, they must know what the test measures and what level of these skills is necessary. They must also be familiar with the abilities of the reference group, as a group. They do not have to identify specific individuals as qualified or not qualified, but they must be able to judge approximately how many are qualified.

Defining adequate and inadequate levels of the knowledge and skills tested can be done in the same way as for the methods we have discussed previously. This is an important step in the process, in this method as in any other method, because this definition will determine the meaning of the standard.

The judges can make their judgments individually or as a group. Again, we recommend a compromise procedure:

1. Have each judge make a preliminary judgment.
2. Write the judgments on a blackboard or a large sheet of paper.

*In 1981, the National Board of Medical Examiners changed from a standard based on current test-takers to a standard based on previous test-takers, for exactly this reason. (*The National Board Examiner*, v. 28, no. 1, Winter 1981. Philadelphia: National Board of Medical Examiners.)

3. Ask for a judge who chose a high number to explain why. Then ask for a judge who chose a low number to explain why. Allow some discussion, but do not try to get all the judges to agree.
4. Give the judges a chance to change their judgments if they want to. Then collect the revised judgments.

You can combine the judgments by computing the mean, the median, or the trimmed mean, as described earlier on pages 22-23.

The main limitation of this method is that the judges must be able to judge the number or the percentage of the test-takers in the reference group who are qualified in the knowledge and skills the test measures. This kind of judgment is not easy to make with any reasonable degree of precision. However, if you can get an approximate judgment of this type, you can use this method as a reality check on the methods based on judgments about test questions. For example, if you can be fairly sure that at least 75 percent of last year's test-takers were qualified, you should be skeptical of any method that produces a passing score that would have passed less than half of last year's test-takers.

One example of setting passing scores by using judgments about groups of test-takers is the awarding of college course credit, on the basis of an examination, to students who have not taken the course. Typically, the college will have the students in the course take the accreditation test at or near the end of the course. When the students' grades have been determined, the testing office computes the distribution of test scores for the A students, for the B students, and so on. The college can then set the passing score on the basis of these distributions. One popular choice is the "mean C"—the average test score of the C students. This choice means that if a student who has not taken the course can score as high on the test as the average C student did *after* taking the course, that student will get credit for the course.

Another method based on judgments of groups of test-takers is similar to the contrasting-groups method described earlier, except that it does not require judgments of individual test-takers. Instead, you identify a group of persons who can be presumed to have the qualifications the test is intended to measure and a group of persons who can be presumed to lack these qualifications (for example, students who have had the relevant instruction and students who have not*). You then select a sample of persons from each group (the same number of persons from each) and give them the test. You set the passing score at the test score level that best discriminates between the two samples. This method will

*See the article by R. A. Berk listed in the bibliography.

not necessarily produce the same result as the contrasting-groups method based on judgments of individual test-takers. Therefore, it will not necessarily minimize the number of wrong decisions in the group of test-takers the test is intended for. It will do so only if (1) the test scores of the "qualified" group are representative of the scores of the qualified people who will be taking the test, and (2) the test scores of the "unqualified" group are representative of the scores of the unqualified people who will be taking the test, and (3) the proportions of "qualified" and "unqualified" people are the same in the standard-setting study as in the group of people the test is intended for.

Choosing a Standard-Setting Method

Which Method is Best?

There is no one method that is best for all testing situations. Your choice of a method should depend on what kind of judgments you can get—and believe. We believe that the best kind of data to use—if you can get them—are the test scores of real test-takers whose performance has been meaningfully judged by qualified judges. If you can have the judges actually observe the test-takers' performance or samples of their work, we recommend the contrasting-groups method. This situation will occur fairly often with essay tests, hands-on performance tests, etc. For multiple-choice tests, we recommend using the contrasting-groups method whenever you can be reasonably sure that the judges will base their judgments on the same qualities of the test-takers—the same knowledge and skills—that the test measures. The contrasting-groups method has the strongest theoretical rationale of any of the methods we have presented: that of statistical decision theory. It is the only standard-setting method that enables you to estimate the frequencies of the two types of decision errors. The main disadvantage of the contrasting-groups method is the difficulty of getting the necessary judgments.

If you cannot get valid judgments of an appropriate sample of the test-takers,* but each judge can confidently identify individual test-takers as good examples of people with "borderline" qualifications, we recommend the borderline-group method. If the judges can best express their standards in terms of the performance of a particular group of test-takers (for example, "at least as good as the average C student"), we recommend setting the standard in those terms.

If none of these conditions can be met, we suggest you use one of the methods based on judgments about test questions—Nedelsky's, Angoff's, or Ebel's—but we also suggest you compare the results of that method with real test-score data. Be prepared to compromise if this comparison suggests that the judges' standards were unrealistic.

Methods such as Nedelsky's, Angoff's, and Ebel's are especially useful when it is important that the passing score represent the standard of

*See pages 35-36 of this manual.

a large and diverse group of people. For example, in choosing the passing score on a math test used as a requirement for high school graduation, it may be important to include the opinions of parents, employers, and community leaders. These people are not in a position to observe the mathematical skills of high school students, so they cannot serve as judges in the borderline-group or contrasting-groups method. But they can serve as judges in Nedelsky's, Angoff's, or Ebel's method.*

Nedelsky's, Angoff's, and Ebel's methods require the judges to review the test. If security considerations prevent you from showing the test even to the judges, you may be able to wait and hold the judging session after the test has been given. If you do not have this option, you may be able to collect the judgments and set the standard on another form of the test (containing different questions measuring the same abilities) if the form to be judged will be statistically equated to the form you will be using. If none of these options is open to you, you will not be able to use one of these methods.

In choosing between Nedelsky's, Angoff's, and Ebel's methods, your main concern should be the type of judgments the judges can make most meaningfully. Angoff's method requires the judges either to think in terms of probabilities (which is difficult for many people) or to imagine a group of borderline test-takers (which may be far removed from the judges' experience). However, Angoff's method is the easiest of the three methods to explain and the fastest to use. Ebel's method enables the judges to take account of the difficulty and the importance of each test question. This feature is especially valuable when the questions on the test differ widely in their importance. Its disadvantages are its slowness and its unsuitability for short tests. Nedelsky's method takes account of the fact that the difficulty of a multiple-choice question depends on just how wrong the wrong answers are. However, Nedelsky's method can be difficult to use when the questions are negatively worded or contain other types of complexities.

*An article by R. M. Jaeger, listed in the bibliography, presents another method of the same general type, developed specifically for tests used as a requirement for high school graduation.

Social and Political Issues

Choosing the passing score on a test often leads to controversy. The controversy may focus on your choice of a method or your selection of judges, or it may focus on any of a number of other issues. You should think about these issues before you begin the process of choosing a passing score. Even if you decide not to take positions on some of these issues, you will be better able to avoid destructive controversies—or to resolve them if they occur—if you have thought about the issues beforehand.

Should You Allow Exceptions to Your Decision Rule?

A common criticism of the use of a passing score is that it fails to allow for exceptions. There may be good reasons for making exceptions to a rule. If you decide not to allow any exceptions, you may be forced to make a decision that is unreasonable under the circumstances. For example, a test-taker may have a particular handicap that results in a lower score than other test-takers with the same level of knowledge would get. If you could anticipate all the possible reasons that would justify an exception, you could write them into the decision rule. Unfortunately, no human being can foresee all the possible circumstances in which a decision rule would be unreasonable.

The problem with allowing exceptions is that once you have made an exception, where do you stop? You may find yourself pressured by people seeking exceptions for reasons you do not consider legitimate. Also, exceptions tend to undermine people's faith in the fairness of your decision procedure. An exception that some people regard as compassion may look to others like favoritism.

One way to deal with this dilemma is to have an established procedure for determining whether an exception should be allowed. You might form a standing committee to approve or deny requests for exceptions. If you find a particular type of special circumstance occurring frequently, you can modify your decision rule to cover it. Each time you modify the decision rule in this way, you will reduce the number of exceptions you will have to deal with in the future.

Should You Allow Test-Takers Who Fail the Test to Take it Again?

In most cases the answer to this question will be "yes." A test-taker may have a "bad day" on the day of the test. If so, the test-taker's score will not represent his or her true level of ability. But if you do allow retakes, should you limit the number of times a person can take the test in an attempt to pass? Should you require persons who fail the test to wait a specified length of time before retaking it? Should you require them to take some sort of instruction before retaking the test? Should you retest them with a different form of the test (that is, one with different questions or problems constructed to measure the same general types of knowledge and skills)?

In most cases, a person who retakes a test should be given a different form of the test each time. Otherwise, the person may become a specialist in the specific problems and questions on the test, without learning the more general knowledge and skills those questions are intended to represent. As long as different forms of the test are available, we prefer not to limit the number of times a person can take the test. No matter how many times the person has failed the test, it is always possible that the person's skills may improve.

Whether to require a waiting period for persons who want to retake the test will depend on your particular testing program. If the testing procedure is expensive and the test-takers are not the ones paying for it, you may want to require a waiting period as an incentive for the test-takers to improve their skills before retaking the test. Another way to make sure the test-takers are adequately prepared is to require failing test-takers to have additional instruction in the knowledge and skills to be tested, before retaking the test.

Should Persons Who Have Passed the Test Ever Have to Take it Again?

There are situations in which such a requirement makes a great deal of sense, particularly where an unqualified person represents a danger to others. For example, airline pilots are required to demonstrate their skills not just once, but every six months as long as they continue to fly. In deciding whether this type of requirement makes sense in your testing situation, you should consider questions like these: Could a person's level of ability decrease over time? What could happen if it did? Is the test changing from year to year, to include new knowledge and skills? What could happen if a person has mastered the old knowledge and skills but not the new?

Should You Establish an "Uncertain" Category?

When you use a test with a single passing score, two kinds of decision errors are possible. An unqualified person may get a score above the passing score; a qualified person may get a score below the passing score. One way to reduce the chances of *both* kinds of errors is to establish an "uncertain" category. For persons in this middle category, you will have to get additional information before making the pass/fail decision. This additional information might be another form of the same test, or a different test, or some other type of evaluation.

To establish an "uncertain" category you will have to choose two critical scores instead of only one, since you are dividing the test-takers into three groups instead of only two. With some methods, this modification will double the time and effort required. However, with the contrasting-groups method, you may be able to choose two critical scores with very little extra work. If you have estimated the relationship between a test-taker's score and the probability that the test-taker will be judged as qualified, you can specify the two critical scores in terms of these probabilities. For example, you might decide to pass any test-taker with more than a 75 percent probability of being qualified, fail any test taker with less than a 25 percent probability, and seek additional information about the rest.

There may be situations in which you cannot get any additional information about the test-takers. If no other information is available and the test-taker cannot even retake the test before a decision must be made, an "uncertain" category may not be of much help.

Should You Use Normative Information in Setting an Absolute Standard?

This is, to some extent, a philosophical issue. Even an absolute standard is ultimately normative. That is, people's judgments of what a person *should* be able to do will always depend to some extent on what people *can* do. However, there is often a gap between what people (for example, students) can do and what other people (for example, instructors) think they should be able to do. We believe that if you are using a method based on judgments about test questions, it makes sense to use normative data as a "reality check." In this case, we suggest that you not share the normative information with the judges until *after* they have made their initial judgments. Then you can let them know how a group of real test-takers performed on the test. If the judges' idea of a "borderline" test-taker is someone whose performance approaches or exceeds that of the average actual test-taker, their standards may be unrealistically high. Even if you are using a method based on judgments about in-

dividual test-takers, it may make sense to use normative information about test-takers who were not judged, as a check on the process of selecting the test-takers and collecting the judgments. If you know that most of the test-takers are qualified, and yet the majority of them have test scores like those of the persons who were judged as "unqualified," you have reason to suspect that something went wrong.

Should You Allow the Standard to Change Over Time?

In many types of testing, continuity of the standard is important. For example, if the test is a requirement for a diploma or a certificate, the meaning of the certificate will change if the standard changes. But if the test is changed from year to year, it may be easier in some years and harder in others. One way to maintain a constant standard is to adjust the passing score to account for the differences in the difficulty of the test. However, such an adjustment may *appear* to be a change in the standard, even though its purpose is to *avoid* a change in the standard. Therefore, the adjustment may cause political problems. Fortunately you can also maintain the standard by adjusting the test scores to compensate for the change in the difficulty of the test and leaving the passing score unchanged. This type of adjustment is called "equating." It is an accepted and widely used technique in educational testing, but it requires certain types of information linking the two forms of the test. For example, the two forms of the test may be designed to have several questions in common, or both forms of the test may be given experimentally to a group of test-takers.*

In other types of testing, it may be desirable to have some flexibility in the standard. Conditions may change over time. Technological advances may change the levels of certain skills required in an occupation. A critical shortage of people in an occupation may make it necessary to lower the standard. Changes in the educational needs of the children in a school district may require a revision in the standard. Even in the absence of such changes, experience with the effects of using a particular standard may indicate that a revision would be desirable. Here again, equating is necessary if the test is changed from year to year, to adjust for the differences in difficulty that may result. Without equating if this year's test is easier than last year's test, you may think you are raising the standard when you are actually lowering it.

---

* For more information on equating, see the chapter by W. H. Angoff cited in the bibliography.

Should You Set Different Passing Scores for Different Groups of Test-Takers?

In some decision-making situations, the test-takers may come from different instructional backgrounds. For example, some of the people taking a test for certification in a profession may have completed a formal training course, while others may have acquired their professional knowledge and skills informally, on the job. The test-takers without formal training may tend to do poorly on the test, but much better in a practical work situation like the one in which they have gained their experience. However, the use of a lower passing score for these test-takers may appear to be a purely political concession, even if it is not intended to be. The best solution is to use a test that measures *only* the knowledge and skills the person actually uses on the job (or comes as close as possible to this ideal). Also, make sure the test is easy to read and free of tricky questions (for example, questions containing wrong answers that are nearly correct). If there are pictures or diagrams on the test, make sure they look like the things they are supposed to represent. If the test has already been made up, you may find you have to delete some questions in order to make it fair.

Helpful Hints

When you choose the passing score on a test you are making explicit the lowest level of performance that will be considered acceptable. Some people may think that the level you have set is absurdly low. Others, particularly those that fall below it, will think that the level is unfairly high. It will be difficult to convince either group that your passing score is appropriate, because there is no purely objective way to set standards. All methods of setting standards depend on some type of subjective judgment at some stage of the process. Critics will be able to argue that those judgments were wrong. You will never be able to prove that your passing score is correct, but there are steps that you can take to increase the probability that your passing score will be accepted.

Be Prepared to Explain Why You Are Using a Passing Score

Even though a passing score may lead to fairer decisions than those made on a case-by-case basis, some people will perceive the use of a passing score as arbitrary and unfair. You should be prepared to explain the reasons for the use of a passing score in your particular testing situation. In particular, you should be prepared to answer the following questions:

How are the decisions to be made on the basis of the passing score being made now?

Why will the use of the passing score be preferable to the current system?

You should try to anticipate any harm that might be caused by the use of a passing score. You should also be ready to point out the harm that would be caused by *not* using a passing score—that is, by making the decisions the way they would be made otherwise.

Evaluate the Test

The test should be adequately reliable and valid for its intended purpose. It should be free of bias; groups of test-takers should not differ systematically in their scores unless they truly differ in the knowledge and skills the test is intended to measure. If the techniques of evaluating test score reliability, validity, and lack of bias are not among your competen-

cies, get help from people who do understand these techniques *as they apply to tests used with passing scores*. An explanation of these techniques is beyond the scope of this manual, but it is obviously impossible to set acceptable standards on unacceptable tests.

In addition to any empirical evidence of test quality, you should obtain judgments about the test from people who represent those who will be affected by the test. Their opinions about the appropriateness of the test will be important in determining their acceptance of the passing score. A dozen favorable references in the *Mental Measurements Yearbook* will not persuade people that your test is acceptable if the test simply does not look right.

Make Sure the Judges Understand What the Test is Supposed to Measure

Some of the methods we have presented (Nedelsky's, Angoff's, Ebel's) require the judges to review the test in detail. Others do not. We recommend that, no matter what method you are using, you have the judges look at the test, unless you have a reason not to (for example, test security). We also suggest that you give the judges a concise description of the knowledge and skills the test is intended to measure. If you cannot allow the judges to look at the actual test, we suggest you give them a *detailed* description of the knowledge and skills the test is intended to measure *and* a few sample questions similar to those on the test. This kind of preparation will help to guard against the kind of misunderstanding that can lead to judgments that are not based on the abilities measured by the test.

Make the Process of Setting the Passing Score as Open as Possible

The fact that a passing score will be set and the way in which it will be set should be well publicized. People should have a chance to provide suggestions and comments early enough in the process to allow you to act on the information that you receive. For example, if you are setting the passing score for a test to be used as a requirement for a high school diploma, parents, students, teachers, school board members, school administrators, members of community groups, and local employers should all be encouraged to participate. In many situations it will be important to involve members of racial, ethnic, and cultural minorities.

Though it may be impossible to have face-to-face meetings with all the people who may be interested, you can encourage them to write to you about their concerns. Make it hard for people to say that they did not have a chance to become involved and state their views. The more

public involvement there is throughout the standard-setting process, the more likely that the passing score will be accepted when the process is completed.

Make Sure People Understand How the Test Will be Used

It is important that people understand why the test is being given. They should know what kinds of decisions will be made on the basis of the test scores and what kinds will not. If a possible use of a test threatens people, and if you do not intend to use the test in that way, say so. For example, if a certification test will be used only for the certification of new applicants, be sure to tell currently certified people that this requirement will not be applied to them. In a school setting, if a test is to be used only to identify students needing remedial work, state explicitly that the scores will not be used to evaluate teacher performance. Try to anticipate people's concerns about threatening but unintended uses of the test and make public guarantees that the test will not be used in those ways.

Give Adequate Notice That a Passing Score Will Be Applied

It is unfair to make people comply with new requirements unless they are given enough time to prepare. In some instances, it may even be illegal because due process of law requires adequate prior notice of a new rule that may deny benefits to a person. Whether or not prior notice is required (and the kind of notice required) will vary with the situation. For example, people may have entered an accredited training program under the condition that they would receive their certification after completing the program and acquiring a certain amount of experience. The imposition of a new barrier, in the form of a test that they must pass, could lead to legal challenge. People who enter the program knowing that they will have to pass a test are far less likely to challenge that requirement. It may be wise to consult a lawyer before you institute a passing score that may be used to deny benefits to people who would otherwise be eligible for them.

Develop Informative Score Reports

We believe that a person who has taken a test that is used with a passing score should receive *at least* two types of information: his or her own score and the passing score. You should also consider providing additional diagnostic information that might be useful to the test-taker. Even though the pass/fail decision may be based on a single total score, the

test-taker may find it useful to know how many questions he or she answered correctly, and how many incorrectly, in each of the main content categories. Few things are more frustrating than receiving a failing grade on a test without being given any other information. The more information you provide, the more likely it is that your testing program will be accepted.

Allow Plenty of Time for Choosing the Passing Score

One good way to plan your schedule is to count backwards from the time the process must be completed. Be sure to allow time for all the necessary activities. You will have to select and train the judges. If you are using a method you have not used before, you should allow time for a small-scale practice run to make sure the procedure will work properly. You may have to allow time for printing the test, administering the test, and scoring the test.

Even if the test is available and you are setting the passing score by a method that does not require test administration, the process may take longer than you anticipate. For example, it may be difficult to find a time when all of the judges are free. You may discover, after collecting the judgments, that some of the judges simply did not understand what they were doing. In this case, you will have to repeat the judging process. In setting a passing score, as in other areas of life, it is usually wise to assume that if something can go wrong, it will.

Review the Process

Before you actually begin to use the passing score to make decisions, review the procedure by which the passing score was chosen. If something in the process was not right, you will want to find out about it before you have begun to apply the passing score. The following questions may help to focus your attention on things that might have gone wrong:

Were the judges all qualified to make the kinds of judgments they were making?

Were the judges a representative group?

Did the judges understand their task?

Did the judges have enough time to complete their task carefully?

Were all the necessary calculations done correctly?

In addition, if the method was based on judgments about test-takers, consider the following questions:

Did the judges know enough about the test-takers to make valid judgments?

Did the judges concentrate on the same knowledge and skills that the test is intended to measure?

Observe the Effects of Using the Passing Score

Once you have begun to use the passing score to make decisions, try to get information that will enable you to judge its appropriateness. Make an effort to get opinions from the different types of people who are affected. In the schools, these would include administrators, teachers, students, and parents. In an occupational setting, these would include the test-takers, their colleagues, and their supervisors. Try to find out what happened to people who failed the test. Is there evidence that many of them were actually qualified at the time they took the test? Is there evidence that many of the people who passed the test were unqualified? What were the consequences of failing a qualified person? Of passing an unqualified person?

The information you get may well be inconclusive. However, it may indicate that the passing score was clearly too high or too low. In that case, you should be prepared to revise it.

Conclusion

All methods of standard setting require judgment. The process of setting a standard can be only as good as the judgments that go into it. The standard will depend on whose judgments are involved in the process. In this sense, all standards are subjective. Yet, once a standard has been set, the decisions based on it can be made objectively. Instead of a separate set of judgments for each test-taker, you will have the same set of judgments applied to all test-takers. Standards cannot be objectively determined, but they can be objectively applied.

Bibliography

Note: This bibliography is limited to works that have been published as of July 1981 and deal with the problem of setting standards.

Andrew, B. J. and Hecht, J. T. "A Preliminary Investigation of Two Procedures for Setting Examination Standards." *Educational and Psychological Measurement*, 1976, v. 36, no. 1, pp. 45-50. (Report of a small-scale experiment comparing Nedelsky's method and Ebel's method.)

Angoff, W. H. Scales, Norms, and Equivalent Scores. In R. L. Thorndike (ed.), *Educational Measurement*. Washington, D.C., American Council on Education, 1971, pp. 514-515. (Source document for Angoff's method.)

Berk, R. A. "Determination of Optimal Cutting Scores in Criterion-Referenced Measurement." *Journal of Experimental Education*, 1976, v. 15, no. 4, pp. 4-9. (A method based on the comparison between instructed and uninstructed students.)

Brennan, R. L. and Lockwood, R. E. "A Comparison of the Nedelsky and Angoff Cutting Score Procedures Using Generalizability Theory." *Applied Psychological Measurement*, 1980, v. 4, no. 2, pp. 219-240.

Bunda, M. A. and Sanders, J. R., eds. *Practices and Problems in Competency-Based Measurement*. Washington, D.C.: National Council on Measurement in Education, 1979, Chapter IV, "Standards," pp. 47-88. Contains articles by R. M. Jaeger, L. A. Shepard, and L. E. Conaway.

Chuang, D. T., Chen, J. J., and Novick M. R. "Theory and Practice for the Use of Cut-Scores for Personnel Decisions." *Journal of Educational Statistics*, 1981, v. 6, No. 2, pp. 129-152. (Mathematical formulas, derivations, and proofs.)

Ebel, R. L. *Essentials of Educational Measurement*. Englewood Cliffs, N.J.: Prentice-Hall, 1972, pp. 492-494. (Source document for Ebel's method.)

Hambleton, R. K. "Test Score Validity and Standard-Setting Methods." In R. A. Berk (ed.), *Criterion-Referenced Measurement: The State of the Art*. Baltimore: Johns Hopkins University Press, 1980, pp. 80-123.

Huynh, H. "Statistical Consideration of Mastery Scores." *Psychometrika*, 1976, v. 41, no. 1, pp. 65-78. (Mathematical theory for the contrasting-groups method.)

Jaeger, R. M. "An Iterative Structured Judgment Process for Establishing Standards on Competency Tests: Theory and Application." *Educational Evaluation and Policy Analysis*, in press.

Journal of Educational Measurement, 1978, v. 15, no. 4. Special issue on standard-setting. Contains articles by G. V. Glass, N. W. Burton, M. Scriven, R. K. Hambleton, J. H. Block, W. J. Popham, R. L. Linn, and H. M. Levin.

Koffler, S. L. "A Comparison of Approaches for Setting Proficiency Standards." *Journal of Educational Measurement*, 1980, v. 17, no. 3, pp. 167-178. (Report of a large-scale experiment comparing Nedelsky's method and the contrasting-groups method.)

Livingston, S. A. "Choosing Minimum Passing Scores by Stochastic Approximation Techniques." *Educational and Psychological Measurement*, 1980, v. 40, no. 4, pp. 859-873. (Includes a detailed presentation of the up-and-down method.)

Livingston, S. A. "Comments on Criterion-Referenced Testing." *Applied Psychological Measurement*, 1980, v. 4, no. 4, pp. 575-581.

Meskauskas, J. A. and Norcini, J. J. "Standard-Setting in Written and Interactive (Oral) Specialty Certification Examinations." *Evaluation and the Health Professions*, 1980, v. 3, no. 3, pp. 321-360.

Nedelsky, L. "Absolute Grading Standards for Objective Tests." *Educational and Psychological Measurement*, 1954, v. 14, no. 1, pp. 3-19. (Source document for Nedelsky's method.)

Popham, W. J. *Modern Educational Measurement*. Englewood Cliffs, N.J.: Prentice-Hall, 1981. (See Chapter 16, "Setting Performance Standards," pp. 371-399.)

Schoon, C. G., Gullion, C. M., and Ferrara, P. "Bayesian Statistics, Credentialing Examinations, and the Determination of Passing Points." *Evaluation and the Health Professions*, 1979, v. 2, no. 2, pp. 181-201.

Shepard, L. "Standard Setting Issues and Methods." *Applied Psychological Measurement*, 1980, v. 4, no. 4, pp. 447-467.

Skakun, E. N. and Kling, S. "Comparability of Methods for Setting Standards." *Journal of Educational Measurement*, 1980, v. 17, no. 3, pp. 229-235. (Report of a small-scale experiment comparing Nedelsky's method and Ebel's method.)

Appendix

Additional Calculations Required by the Correction for Guessing

The usual correction-for-guessing formula used with multiple-choice tests depends on the number of choices per question. If each question has five answer choices, four of them will be wrong answers. The traditional correction for guessing is to subtract one-fourth of the number of wrong answers the test-taker chooses. Similarly, if each question has four answer choices, three of them will be wrong answers; to correct for guessing, subtract one-third of the number of wrong answers the test-taker chooses.

The Nedelsky, Angoff, and Ebel methods produce an estimate of the expected number of correct answers the "borderline" test-taker will choose. To find the test-taker's expected score, corrected for guessing, do the following calculations:

1. Subtract the expected number of correct answers from the total number of questions to get the expected number of wrong answers.
2. Divide this number by the number of *wrong* answers per question, to get the expected number of penalty points.
3. Subtract this number from the expected number of right answers, to get the test-taker's expected score.

For example, suppose the test has ten questions and each question has five answer choices, as in Table 1 on page 22. If the expected number of correct answers is 4.43, you would do the following calculations:

Expected number of wrong answers: $10 - 4.43 = 5.57$

Expected number of penalty points for guessing: $5.57 \div 4 = 1.39$

Expected score, corrected for guessing: $4.43 - 1.39 = 3.04$

51055-00561 • RR99P2.5 • Printed In U.S.A.
I.N. 239660