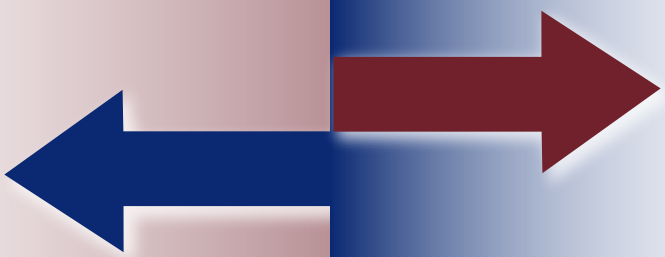
A Primer on Setting Cut Scores on Tests of Educational Achievement

Michael Zieky & Marianne Perie

Including

Excerpts From *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*

Samuel Livingston & Michael Zieky



Listening.
Learning.
Leading.

Foreword

A Primer on Setting Cut Scores on Tests of Educational Achievement is timely and useful given the greater emphasis today on using test scores to make important decisions. It is part of ETS's mission to help advance quality and equity in education through the organization's research, and this publication advances this goal by providing practitioners a guide on what to consider when setting standards of proficiency on tests of educational achievement.

Marianne Perie and Mike Zieky note that while there is no perfect way to set cut scores, there are certain steps to follow and established methods to use that will produce sensible and useful cut scores. The authors also emphasize that cut scores must be validated and that practitioners should be prepared to make changes if experience shows that the cut scores are not meeting their intended purpose.

Regards,

Ida Lawrence
Senior Vice President
Research & Development
ETS

Purpose

We created this primer for educators and policymakers who need a basic understanding of the issues that must be faced and the decisions that must be made in setting cut scores on widely used tests of educational achievement, such as district and state assessments. No knowledge of measurement or statistics is required of readers.

Overview

Cut scores are selected points on the score scale of a test. The points are used to determine whether a particular test score is sufficient for some purpose. For example, student performance on a test may be classified into one of several categories such as basic, proficient, or advanced on the basis of cut scores.

The setting of cut scores on widely used tests in educational contexts requires the involvement of policymakers, educators, measurement professionals, and others in a multi-stage, judgmental process. Cut scores should be based on a generally accepted methodology and reflect the judgments of qualified people.

The primer has three main sections:

The major steps that must be followed to set reasonable cut scores, including:

- determining if cut scores will be useful
- appointing staff for the tasks involved
- selecting the performance levels to be reported (e.g., basic, proficient, advanced)
- describing what students need to be able to do to reach each performance level
- setting provisional cut scores
- establishing operational cut scores
- documenting the process
- evaluating the results of using the cut scores

Important issues to consider in setting and using cut scores, including the:

- necessity of making judgments when setting cut scores
- qualifications of judges involved in various stages of the process
- concept of borderline performance
- likelihood of errors of classification when using cut scores
- reliability of the classifications made on the basis of cut scores
- need to consider aligning cut scores across grades
- choice between compensatory and conjunctive scoring
- importance of normative information

Methods for setting cut scores,[1] including:

- Nedelsky's method
- Angoff's method
- Ebel's method
- the Borderline Group method
- the Contrasting Groups method
- an extension of Angoff's method
- the Bookmark method
- the Body of Work method

[1] With the exception of the discussions of newer methods (extended Angoff, Bookmark, and Body of Work), the material on methods for setting cut scores has been excerpted from *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests* by Samuel A. Livingston & Michael J. Zieky, originally published by ETS in 1982 and reprinted in 2004.

Determining If Cut Scores Will Be Useful

The first step is for policymakers to specify exactly why cut scores are being set in the first place. The policymakers should describe the benefits that are expected from the use of cut scores. What decisions will be made on the basis of the cut scores? How are those decisions being made now in the absence of cut scores? What reasons are there to believe that cut scores will result in better decisions? What are the expected benefits of the improved decisions?

It is important to list the reasons why cut scores are being set and to obtain consensus among stakeholders that the reasons are appropriate. An extremely useful exercise is to attempt to describe exactly how the cut scores will bring about each of the desired outcomes. It may be the case that some of the expected benefits of cut scores are unlikely to be achieved unless major educational reforms are accomplished. It will become apparent that cut scores, by themselves, have very little power to improve education. Simply measuring a child and classifying the child's growth as adequate or inadequate will not help the child grow.

For example, if one of the reasons for setting cut scores is to improve the quality of instruction in the schools, it should become clear that cut scores by themselves would not have the desired effect. The use of cut scores may point out that certain schools, curricular areas, demographic groups, regions, and so forth are more in need of improvement than others, but the cut scores alone will not improve education. Unless the infrastructure needed to improve education is put in place, setting cut scores will be futile for that purpose.

It is also necessary to consider the potential negative effects of setting cut scores. What will happen to students who fail? Will they be stigmatized and ignored or will they be helped? What will happen to schools with large proportions of failing students? Will the institutions be punished or assisted? What will happen to teachers with large numbers of failing students? Will the teachers be punished or will they receive additional help?

Even though the use of cut scores may lead to positive consequences, some people will perceive the use of cut scores to be unfair. It is important to explain the reasons for the use of cut scores to educators and to the public. People should understand why the tests are being given and why students are being classified into different proficiency levels.

Appointing Staff for the Tasks Involved

If there are generally agreed-upon reasons for setting cut scores, and the expected benefits clearly outweigh the expected negative consequences, it is reasonable to go on to the next step. The second step is for the policymakers to appoint or hire managers for the process of setting cut scores. A common practice is for the policymakers to release a request for proposals for the task of setting cut scores. The managers should include experts in setting cut scores. The experts may be found among local educators, but often they are external consultants or on the staff of testing agencies contracted to help set the cut scores.

The managers, within the constraints established by the policymakers, will determine the tasks that must be accomplished, establish schedules and budgets, select the participants in the various stages of the process, determine the methods used to set the cut scores, train the judges, monitor progress, ensure that the necessary logistics are managed, document results, and take on responsibility for control of the entire process.

Selecting the Performance Levels to Be Reported

The policymakers, with the help of educators, should decide the performance levels to be reported and the general definitions of each level. Performance levels indicate the categories into which student performance will be classified. Performance levels are stated in general terms that can be applied across the subject areas and grades to be tested. For example, policymakers may decide to use performance levels such as pass and fail, or basic, proficient, and advanced. They may define *proficient* as the National Assessment of Educational Progress (NAEP) does, using general terms such as "competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter."

The general performance levels do not depend on a particular method for setting cut scores and may be determined before a method is selected. Essentially, the process is one of obtaining consensus on the number of categories to be used to classify students, the labels to be used for each category, and a general definition of what is meant by each label.

Having no more than three or four categories is best because it may become difficult to differentiate among more of them. NAEP, for example, uses the following proficiency levels:

- The basic level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
- The proficient level represents solid academic performance and demonstrated competence over challenging subject matter.
- The advanced level signifies superior performance.

Describing What Students Need to Know and Be Able to Do to Reach Each Performance Level

The next step is for groups of educators familiar with students in the affected grades and familiar with the subject matter to describe what students should know and be able to do to reach the selected performance levels. Performance level descriptors detail the knowledge, skills, and abilities to be demonstrated by students who have achieved a particular performance level within a particular subject area. In K-12 educational contexts, performance level descriptors are usually tied to grade levels.

Performance level descriptors describe what a student must be able to do to reach each performance level. For example, what does a third-grade student have to be able to do to be considered proficient in reading? What does a fifth-grade student have to be able to do to be considered advanced in math?

The performance level descriptors are best developed by educators who are aware of what students should be able to do based on the instruction they have received in the subject at various grade levels. Performance level descriptors should be approved by the policymakers before they are used operationally.

Writing performance level descriptors can be a big job because separate descriptions are needed for each performance level, in each grade, and in each subject area of interest. For example, if the performance levels basic, proficient, and advanced are selected, and if they are to be applied in Grades 3–8, in language arts and mathematics, then 36 performance level descriptions will be required (3 performance levels x 6 grades x 2 subject areas).

Examples of performance level descriptors in reading for Grade 4 from the National Assessment of Educational Progress	
Basic	Fourth-grade students performing at the basic level should demonstrate an understanding of the overall meaning of what they read. When reading text appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences and extend the ideas in the text by making simple inferences.
Proficient	Fourth-grade students performing at the proficient level should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to fourth grade, they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear.
Advanced	Fourth-grade students performing at the advanced level should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. When reading text appropriate to fourth grade, they should be able to judge text critically and, in general, give thorough answers that indicate careful thought.

In addition to writing performance level descriptors that are individually appropriate for a particular grade and performance level, it is important to write descriptions

that make sense across grades and performance levels. For example, it would be confusing and inappropriate if more knowledge and skill were required to be considered basic in fourth-grade reading than to be considered proficient in fifth-grade reading. Therefore, communication across the groups drafting the performance level descriptions is necessary.

The setting of performance level descriptors has to be completed before the setting of cut scores can begin. Writing performance-level descriptions and setting cut scores at the same meetings using the same people is possible, but sufficient time must be allotted for both tasks.

Setting Provisional Cut Scores

The next step is to set provisional cut scores. A provisional cut score is a preliminary selection of the score needed to meet a given performance level descriptor on a particular test. Cut scores on tests of educational achievement are best determined by people who are aware of what students should know and be able to do based on the instruction the students have received in the subject.

The people who set the provisional cut scores may be, but need not necessarily be, the same people who develop the performance level descriptors. While the performance level descriptors focus on *what* students should know and be able to do, cut scores focus on *how many* score points students have to earn to demonstrate they have reached the level of knowledge and skill indicated by a specific performance level descriptor. Cut scores on academic tests are usually set by educators using any of several procedures that involve judgments about students or judgments about test questions.

The managers, in consultation with the policymakers, should determine which method or methods will be used to set the cut scores. The excerpts from *Passing Scores* and the additional materials in the section on methods describe eight useful ways to set cut scores and provide some information on how to select an appropriate method. Other methods are available in addition to those discussed in this booklet. There are no perfect methods, but some are better than others in particular circumstances.

Once a method is selected, the managers should determine the types and numbers of judges required, the types and number of meetings required, the kinds of data the judges will need, the schedule for accomplishing the tasks, the resources required, and the numbers and kinds of people needed to accomplish such tasks as selecting judges, setting up and running the meetings of the judges, performing computations, and documenting the studies.

There are many open questions about how the meetings of the judges should be run, even after an appropriate method has been selected. There is general agreement, however, that training of the judges in the method that they will apply is important. Although the training will differ depending on the method selected, some basic principles remain the same. The judges should learn about the procedures they will follow, the cognitive task they will perform, and the rationale behind the task. The judges should be told how their individual judgments will be combined to produce a single cut score. If possible, there should be practice in the process of making the judgments on sample questions. After the judges have had a chance to practice, the trainer should employ some methodology, such as an evaluation form, to ensure that the judges understand the process, are comfortable with their task, and are ready to proceed.

The questions to be considered by the managers include:

- How many judges and what qualifications and experience should they have?
- How can judges be selected to ensure that they reflect a cross section of important experience and demographic variables?
- What normative information should be given to the judges and when should they receive it?
- When should the discussions of the judgments take place and how extensive should they be?
- How many iterations of judgments should there be?
- When should judges be told the likely effects of their judgments on the cut score indicated by the study?
- Should any judgments be excluded from the calculation of the provisional cut score? If so, what criteria should be used to exclude judgments?
- How close a linkage is required across grades and how will the linkage be achieved?

- Should compromise methods be used that combine normative and absolute judgments? Which method is most appropriate?

The task of setting cut scores can be large. Research suggests that 10–15 judges should be used for a group setting cut scores. For example, assuming cut scores are to be set on tests of reading and math for Grades 3–8, with a separate group for each subject area within each grade, 12 group meetings will be needed. The meetings are likely to take at least two full days each. If, on average, 13 judges are at each meeting, the meetings will consume 312 judge-days (13 judges x 12 meetings x 2 days per meeting). In addition to the judges, each meeting will require a lead person who is an expert in standard setting, a content specialist, a person to perform the necessary computations, and one or two support staff. If four staff members run each meeting, the meetings will consume 96 staff-days (4 staff x 12 meetings x 2 days per meeting). Additional staff time will be needed to document the meetings and the results.

The time required can increase because separate cut scores may need to be set on the same test if it will be used for several purposes. For example, a university admitting non-native speakers of English to graduate school may have different requirements for fluency in English for students enrolling in the English department than for students enrolling in the mathematics department. A single compromise cut score may be inappropriate for both departments.

The provisional cut scores are based on subject matter-content considerations. They may be modified based on policy considerations and should be adopted or approved by the policymakers before operational use.

Establishing the Operational Cut Scores

The next step is for policymakers with the authority to do so to establish the operational cut scores, which are the cut scores actually used to classify students into various performance levels. The numbers that result from the cut score meetings are provisional. The cut scores may appropriately be modified for policy considerations. For example, policymakers may decide that in marginal cases it is preferable to pass than to fail a student. They may, therefore, adjust the cut score accordingly. They may adjust cut scores to better align the passing rates in different grades. Because every test has some random fluctuations in scores, the policymakers may adjust the cut scores to reduce the likelihood that any student would fail (or pass) because of those random fluctuations.

The job of the policymakers at this stage is essentially to make the adjustments that increase the likelihood that the cut scores will help to support the purpose of the assessment program.

Questions such as the following should be considered before cut scores are released for operational use:

- Were all the judges 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 tend to agree with each other?
- Did the judges have enough time to complete their task carefully?
- Were all the data entered correctly?
- Were all the necessary calculations done correctly?

The policymakers should also consider such questions as:

- Is it better to pass a student who deserves to fail, or fail a student who deserves to pass?
- How should variability in the judges' ratings affect the revision of a provisional cut score?
- Do any revisions need to be made for the sake of improved linkage across grades or across subjects?
- Do the results make sense given what is known about the students, the schools, and the curriculum?
- Do the results further the aims of the testing program?

Documenting the Process

The managers should document all aspects of each provisional cut score study and the policy-level revisions that resulted in each operational cut score including the:

- rationale for the selected method
- qualifications of the people who ran the study
- qualifications and demographic descriptions of the judges
- training received by the judges
- steps followed during the study
- data that were shared with the judges
- judges' ratings at every stage at which ratings were made

- evaluation forms completed by the judges
- rationales for any adjustments made by policymakers

The best strategy for documentation is to act as though someone will be called on to reconstruct and defend every aspect of the cut score study long after the participants have forgotten the details of what occurred.

Evaluating the Results of Using the Cut Scores

The final steps are evaluative and should be applied after the cut scores have been in place for at least one round of testing and reporting of results. The validation of cut scores requires a systematic effort to determine whether the cut scores are working appropriately. Are the results reasonable? Are failure rates far higher or lower than expected? Are even outstanding students failing to reach the highest level? Or, on the other hand, are poorly prepared students reaching the highest level? In general, do the performance levels in which students fall comport with the grades they receive in school?

Is the use of cut scores having the desired effects? One of the earliest steps of the process was to define the expected benefits of using cut scores. What evidence is there that the benefits are being achieved? What negative consequences have arisen? Is it clear that the benefits outweigh the negative consequences? If not, what changes should be made?

Ideally, additional validation studies should be carried out periodically to help ensure that the benefits of using cut scores continue to outweigh the negative consequences.

Important Issues

Necessity of Making Judgments When Setting Cut Scores

All procedures for setting cut scores require the application of judgment. For example, some types of cut score studies require judges to estimate the probability that a hypothetical group of students would know the answer to a test question. Another type of study requires judges to examine a student's performance and to decide whether the performance is good enough for some particular purpose.

No purely objective methods exist. There are no "true" cut scores that a group of perfectly selected, perfectly trained judges using a perfect method will find. The cut scores, rather, reflect the combined judgments of the people involved.

Qualifications of Judges Involved in Various Stages of the Process

People involved at different stages of the process of setting cut scores should have somewhat different characteristics. Policymakers must be aware of the considerations that drive the assessment system. They should be aware of what the assessment and the cut scores are intended to accomplish. They must be aware of the effects of the operational cut scores on students, faculties, schools, and public opinion.

The judges who set cut scores in educational contexts must include educators who are subject-matter experts and who are familiar with students. They should know what is taught to students at the relevant grade levels, and they should be aware of what students actually learn. For high school graduation tests, the judges may also include representatives of institutions of higher education that admit the graduates, and representatives of industries that hire the graduates.

The judgments of equally intelligent, equally experienced, equally well-trained, equally knowledgeable, and equally well-meaning people will differ. Different groups of judges are likely to set different cut scores, because different people have different opinions and cut scores are an expression of the judges' opinions. There is no way to avoid this.

Therefore, the samples of people who will participate in the process of setting the cut scores must be selected with great care. The people at all stages of the process should represent the relevant constituencies. The gender, ethnic, and racial composition of the group should represent the population of possible judges. In addition, if different types of schools are affected by the assessment, judges should represent the different types of schools. If regional differences exist in education, judges should represent the different regions. In short, the judges who participate should be as representative as possible of all the people who could have been used as judges.

Concept of Borderline Performance

The concept of borderline performance is very important in setting cut scores. For example, to determine which students are proficient, it is necessary to distinguish students who are basic from those who are proficient, and students who are proficient from those who are advanced. Therefore, distinguishing between the best performing student who is still basic and the worst performing student who is still proficient is neccessary. Similarly, distinguishing between the best performing student who is still proficient and the worst performing student who is still advanced is neccessary.

Focusing on the average student within a category will not help to make the necessary distinctions. The focus should be at the point at which the best performing student within a category becomes indistinguishable from the worst performing student in the next higher category, or the point at which the worst performing student in a category becomes indistinguishable from the best performing student in the next lower category. In pass-fail terms, judgments have to focus on the borderline between the best performing student who still deserves to fail and the worst performing student who still deserves to pass.

Likelihood of Errors of Classification When Using Cut Scores

Two types of errors are likely to occur when cut scores on tests are used to classify students. These errors of classification do not occur because someone made a mistake. The errors will occur because no test can be perfectly reliable or perfectly valid, and because no method of setting cut scores is perfect. In the following discussion, the categories of pass and fail will be used because they are easiest to understand in this context. The same logic applies, however, regardless of the labels chosen for the performance levels, or the number of levels chosen.

If the cut scores are set too high, then students who really deserve to pass will fail. If the cut scores are set too low, then students who really deserve to fail will pass. Moving the cut score up or down to reduce one type of error will necessarily increase the chances of making the other type of error. For example, it is possible to reduce the number of students who pass, but who really deserve to fail, by raising the cut score. The cost of doing so, however, is to increase the number of students who fail but who really deserve to pass. Good test development and good practices for setting cut scores can reduce the number of errors of classification, but no way exists to reduce the errors to zero.

Sometimes the relative harm caused by the two types of errors is easy to determine. For example, if a test is used to license airline pilots, passing test takers who deserve to fail is clearly more harmful than failing test takers who deserve to pass. In most academic settings, however, the relative harm caused by the two types of errors is much more difficult to determine.

For example, some people say that using rigorous standards that fail marginal students merely punishes students for failures of the educational system and that failing the students does much more harm than good. Other people say that the application of rigorous standards is the only way to improve schools and that passing marginal students who may lack important skills is harmful to the students and to society. Which group is right? The decision is a matter of values, not of empirical truth.

People with one set of values will attack a cut score as being too high while people with a different set of values will attack the same cut score as being too low. There is no way to determine which group is right, because the

correctness of the decision depends on judgments about which type of error of classification is worse. Reasonable people will disagree.

The people involved in setting operational cut scores should consider both types of errors in making their judgments and decide which type of error they consider more harmful. The cut scores should reduce the more harmful type of error. Note that if one of the types of errors causes no harm, no need exists to set cut scores. For example, if passing students who deserve to fail causes no harm at all, the best strategy is simply to pass all students.

Reliability of the Classifications Made on the Basis of Cut Scores

A useful synonym for "reliability" in an assessment context is "consistency." If the student stays the same, how consistent will her scores be if she takes an alternate form of the test (different questions covering the same content at the same difficulty)? For tests that are used with cut scores, it is important to get answers to the following questions:

- What proportion of students would be classified the same way if they had taken a different form of the same test?
- What proportion of students would be classified the same way if they had taken the same form on a different day (assuming no changes in knowledge)?
- What proportion of students would be classified the same way if their responses to the constructed-response questions, such as essays, had been scored by different people?

All of those questions deal with the reliability of the classifications being made. The reliability of classification will not be perfect, even for good tests. Every test is only a sample of all the questions that could be asked. Test takers are not likely to be equally knowledgeable about all of the legitimate questions that could be asked, so test form to test form differences are likely. Day-to-day fluctuations in students' attention, memory, luck in guessing, and so forth are also expected.

Even scorers who have been well trained will disagree occasionally about papers that are near the borderlines of score differences. For example, a response that one judge considers a high 2, another judge may legitimately see as a low 3. This discrepancy becomes a problem if the cut score requires a minimum of 3 to be classified as proficient. The discrepancy would not affect the reliability of classification, however, if scores of both 2 and 3 were considered basic.

Students with similar scores on a test tend to be similar in what they know about the tested subject. Most tests cannot distinguish well between students with scores that are very close to one another. Whenever a cut score is used, however, students with scores just above the cut score and students with scores just below the cut score will be classified differently. What this means is that students who score near the cut score may pass or fail a test because of random fluctuations.

The more reliable a test is, however, the less likely it is that the scores will be affected by large random fluctuations. All other things being equal, longer tests will be more reliable than shorter tests; and objectively scored tests will be more reliable than subjectively scored tests. Note that if the reliability of classification is such that test takers are classified the same way 50% of the time, the same level of consistency could be achieved by flipping a coin. At that level of reliability, the test is providing no useful information about the proficiency levels of individual students.

Need to Consider Aligning Cut Scores Across Grades

When cut scores are being set in different grades by different groups, it is necessary to build in some form of linkage across the grades to avoid anomalies such as requiring more knowledge and skill from a proficient sixth-grade student than from an advanced seventh-grade student. Lack of linkage across grades may result in inappropriate cut scores, misclassify students, and provide misleading data to educators and the public. It may be appropriate, however, to require more knowledge and skill from an advanced sixth-grade student than from a basic seventh-grade student. That is a matter of judgment for people who are familiar with the curriculum and the proficiency levels.

At higher grade levels, performance in some subject areas may stop changing across grades. Reading, for example, is likely to have large differences between grades in the early grades, but have small or non-existent differences in the higher grades. For example, third-grade students are likely to read much better than first-grade students do. Twelfth-grade students, however, are not likely to read much better than 10th-grade students do. In other areas, such as speaking a foreign language, growth is likely to continue even in the higher grades. Linkage does not require equal increments in knowledge and skill across grades and subject areas. The goal is to avoid inappropriate reversals and widely discrepant, unjustified differences in passing rates from grade to grade.

If tests are vertically scaled (a single score scale is used across all of the tested grades with the expectation that students in the lower grades would score lower on the scale than would students in the higher grades), then the cut scores should increase as the grade levels increase. For example, if the math test is vertically scaled, it would be inappropriate for the proficient cut score in Grade 5 math to be 615 and the proficient cut score in Grade 6 math to be 575.

One straightforward way to link cut scores across grades is for the policymakers to adjust the provisional cut scores to reduce reversals in vertically scaled cut scores and unjustified grade-to-grade differences in passing rates.

Choice Between Compensatory and Conjunctive Scoring

An important issue in using cut scores when there is more than one score is whether the scoring should be compensatory or conjunctive. In conjunctive scoring, each separately scored content area has to be passed at some separate cut score. Scores on one content area have no effect on whether a student passes a different content area.

In compensatory scoring, high scores on one content area can compensate for low scores on another content area. Consider a test that provides separate scores for algebra and geometry. If the cut score on geometry is 26 and the cut score on algebra is 30, a student who scored 20 on geometry and 36 on algebra would fail if conjunctive scoring were required. However, the student would pass if compensation were allowed.

Which type of scoring is preferable? Because longer collections of test questions tend to be more reliable than shorter collections of test questions, compensatory scoring tends to be more reliable than conjunctive scoring. In conjunctive scoring, if a student has to pass all of the content areas separately, the least reliable score controls whether a student will pass.

For some subject areas, compensation does not make sense. As a clear example, tests of English for English language learners often measure reading, writing, listening, and speaking. Knowing how to read English really well should not compensate for the lack of ability to speak English. Knowing how to write English really well should not compensate for the lack of ability to comprehend spoken English.

In many academic subject areas, however, compensatory scoring makes a great deal of sense. People are not equally good at everything, and strengths in one area often compensate for weaknesses in other areas.

The Importance of Normative Information

Some people think that the use of normative data will somehow contaminate the cut score process. All cut scores, however, are ultimately based on norms. Judgments of what a person *should* be able to do will always depend to some extent on what judges think people *can* do. For example, nobody would set a cut score that required students to run a mile in three minutes to be classified as proficient in physical education.

Cut scores depend on judgments that may be difficult to make. Therefore, using normative data as a "reality check" makes sense. We think that it is preferable to give all judges good information about what students can actually do rather than to depend on whatever each judge happens to think that students can do. One type of information is the difficulties of the test questions for some defined group of test takers. The difficulties of the questions are usually expressed as the average percent correct on multiple-choice questions or the average score on constructed-response questions. That the average test taker is rarely the borderline test taker whom they should be considering must be made clear to the judges. Nonetheless, real data on the performance of actual students can help the judges make more realistic judgments about the performance

of borderline students. Other data that may be shared with the judges are the pass rates that would have ensued if the tentative cut score that resulted from the judges' deliberations had been used. A common practice is to have the judges make their initial judgments in the absence of data, then to share the results of the first round of judgments and discuss normative data with the judges, and then to have another round of judgments.

Even people who agree that normative information should be given to judges disagree about how to do it. Experts in setting cut scores should be consulted to determine what data should be shared for a particular method of setting cut scores, when the sharing of information is most appropriate, and how many iterations of judgments should be completed.

Methods for Setting Cut Scores

Methods Based on Judgments about Test Questions

(Excerpted from *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*)

The standard-setting methods described in this section 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 earliest article describing one of these methods [Nedelsky, 1954] referred to this person as the "F-D student.") 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 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 cut score in some other way.

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. Choose a cut score by combining the judgments.

The first two steps are the same for all methods (and have been described in earlier sections of this primer). The remaining steps differ.

The three methods we will describe first are named for the people who initially 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.)

One important issue in the application of Nedelsky's method (and of Angoff's and Ebel's methods) 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 cut 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 Cut 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. 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 for a particular question, the individual has one chance in three of choosing the correct answer. Therefore, this person's expected score for that question is 1 divided by 3, or .33.

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 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 cut 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. (One fairly common practice is to eliminate the highest 25% and the lowest 25% of the scores and average the middle 50%. The resulting statistic is called the "midmean.") 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 cut score from their judgments. Otherwise, the judges with the highest and lowest standards may suspect that you are discriminating against them.

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 cut score, a borderline test taker should have a 50% 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,[2] is similar to Nedelsky's method, but it can be used with tests that are not multiple-choice. In Angoff's method, the cut score is computed from the expected scores for the individual questions, as in Nedelsky's method. However,

[2] Angoff attributed the method to Ledyard Tucker.

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 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. Display these numbers so all the judges can see them. If the numbers are all similar (e.g., within 10 or so percentage points), go on to the next question.
   b. If the numbers are not 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.

Choosing the Cut Score.

Finding the expected test score for a borderline test taker is done in basically the same way as in Nedelsky's method. The probability of a correct answer is the test taker's expected score for that question. Simply sum the probabilities for the individual multiple-choice questions to get each judge's estimate of the borderline test taker's expected score for the whole test. 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.

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. (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?"

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 Cut 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 12 categories to get the expected score for the whole test.

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.

Methods Based on Judgments about Individual Test Takers *(Excerpted from Passing Scores)*

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 and the Contrasting Groups method. The main advantage of these methods is that they depend on actual test takers or the products of their work rather than on a hypothetical group of test takers. Furthermore, 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 about the test takers' relevant knowledge and skills.

The first requirement applies to any method of choosing a cut 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 cut 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).

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 cut 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 cut 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 cut 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 cut 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[3] — 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

[3] The method can be used with any contrasting groups such as basic and proficient or proficient and advanced.

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 cut 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 cut score might possibly be located. For example, on a 100-question test, you might choose 10 test takers from each 5-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 cut 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. 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." 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 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, the "smoothed" percent-

qualified for test-score level 86 would be the average of the percentages for test-score levels 85, 86 and 87.

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 adjacent score levels and computing the percent-qualified for this enlarged group. 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 cut scores.

Different smoothing methods can result in different cut 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 cut score.

Choosing the Cut Score.

The final step in applying the Contrasting Groups method is the choice of the cut score. One logical choice is the test score for which the "smoothed" percent-qualified is exactly 50%. 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.

The rationale for setting the cut score at the test score that corresponds to a 50% 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 cut 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 cut 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 cut 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%-qualified. By similar reasoning, if it were three times as bad to pass an unqualified test taker as to fail a qualified test taker, the cut score would be the test score at which qualified test takers outnumber unqualified test takers by three to one. That is, the cut score would be the test score that corresponds to 75%-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 cut 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 cut 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 cut score.

Some Newer Methods of Setting Cut Scores

This section describes three methods of setting cut scores that were developed or refined after *Passing Scores* was published. The newer methods discussed in this section are:

- an extension of the Angoff method
- the Bookmark method
- the Body of Work method

The newer methods of setting cut scores are not necessarily better than the older methods in all situations. An expert in setting cut scores will be helpful in determining which method to use under specific circumstances.

Extension of the Angoff Method

The extension of the Angoff method can be applied to constructed-response questions, such as essays. When judges estimate the probability that borderline test takers will answer a multiple-choice question correctly in the traditional Angoff method, they are actually estimating the expected average score of the borderline test takers for the question. Therefore, a straightforward extension of Angoff's method for constructed-response questions is to have the judges estimate the average score that borderline test takers would obtain on each constructed response question. For example, if the question is worth 6 points, some judges may estimate that the average score of borderline test takers will be 2. Other judges may estimate that the average score of borderline test takers will be 3, just as judges' estimates differ for multiple-choice questions. The estimates of the judges are averaged by using the mean, median, or trimmed mean, as is done for multiple-choice questions. The method can be combined with the traditional Angoff method if a test has both multiple-choice and constructed-response questions. The extended Angoff method shares the advantages and disadvantages of the traditional Angoff method and should be applied following the steps described in the excerpts from *Passing Scores*.

The Bookmark Method

The Bookmark method is used with tests that have been scored using item response theory (IRT).[4] The Bookmark method, introduced in the mid 1990s by Lewis, Mitzel, and Green, has become very popular among state assessment programs. The Bookmark method requires placing all of the questions in a test in order by difficulty from easiest to hardest. Judges are asked to place a "bookmark" at the point between the hardest question borderline test takers would be likely to answer correctly[5] and the easiest question the borderline test takers would not be likely to answer correctly.

Constructed-response questions are included once in the ordered booklet of test questions for each point it is possible to score on the question. That is, a constructed-response question with possible scores of 1, 2, and 3 would appear in the ordered test booklet three times at the difficulty associated with obtaining each score. (This makes sense because it is harder to score 3 points than it is to score 2 points, and harder to score 2 points than it is to score 1 point.)

An "item map" accompanies the ordered test booklet. The "map" includes information on the content measured by each question, information about each question's difficulty, the correct answer for each question, and where each question was located in the test booklet before the questions were reordered by difficulty.

Collecting the Judgments.

The judges are asked to start with the first question and ask themselves if a borderline test taker is likely to answer that question correctly. If the answer is "yes" they proceed to the next question. At the first question they answer "no" they place their bookmark. The facilitator should instruct the judges to continue further into the booklet after the bookmark to ensure that they would also say "no" to the questions following the bookmark. Also, if more than one cut score is being set, the judge would continue through the booklet asking the same question for the borderline student in the next higher level.

Judges should discuss why a borderline student would or would not be able to answer specific questions correctly or reach a particular score on the constructed-response questions. They should also discuss their bookmark locations with the other judges. The judges should have the opportunity to adjust their bookmarks after the discussion. Applications of the Bookmark method generally involve the sharing of data on the difficulty of the questions and on the effects of placing the bookmark at various places. There are generally iterations of the judgmental process after the data have been shared.

Choosing the Cut Score.

Using IRT, measurement experts can translate the location of each judge's bookmark to a cut score on the reporting

[4] IRT is based on the relationships among the difficulty of a test question, a test taker's ability, and the likelihood that the test taker will answer the question (called an "item") correctly. For example, the easier a question is, the more likely it is that a test taker will answer it correctly. Also, the more able a test taker is, the more likely it is that the test taker will answer the question correctly. IRT quantifies the relationships using a mathematical model. IRT allows measurement experts to estimate the score that a test taker of known ability would obtain on a test.
[5] The developers of the method defined "likely to answer correctly" as a probability of .67 or higher of a correct answer.

scale for the test. As with the other methods, the judges' cut scores can be combined using the mean, median, or trimmed mean to obtain a group cut score, although, typically, the median is used for the Bookmark method.

The Bookmark method requires less data entry than other question-judgment methods, takes less time for judges to complete their tasks, and can be used with all types of questions.

Body of Work Method

One type of contrasting groups methodology recently applied to K–12 testing is the Body of Work method. It was described in the early 1990s by Kahl, Crockett, DePascale, and Rindfleisch. This method is used when there are products of test takers' work that can be evaluated, such as essays. The method is difficult to apply with multiple-choice questions alone. The Body of Work method is a holistic procedure in which the judges review all of the test questions and all of a student's responses to the questions.

All of a student's responses to the test questions are placed in what is called a "response booklet." Judges examine an entire student response booklet and match the knowledge and skill demonstrated in the responses to a performance level. The Body of Work method usually includes three rounds of judgments: a training exercise, a range-finding stage, and a pinpointing stage.

The judges begin with the training exercise in which they examine 5–8 response booklets and match them to performance levels. They then share their results with the other judges and discuss the outcomes. The judges are not told the scores of the booklets until the discussion period.

After completing the training exercise, judges typically review 30 or so response booklets in a range-finding round, again categorizing each booklet into its performance level without knowledge of the scores. The response booklets used should represent the range of possible scores. Judges discuss their ratings and re-categorize the papers as they desire.

The results of the range-finding round can be used to determine roughly where each cut score will be by an examination of the distributions of scores of the response booklets placed in each performance level. The task is to find the scores or score ranges that best separate the performance levels. For example, consider the following results:

- All booklets scoring under 46 points were placed in the basic level by all judges.
- Booklets scoring 46–60 points were placed in the basic level by some judges *and* in the proficient level by other judges.
- Booklets scoring 61–71 points were placed in the proficient level by all judges.
- Booklets scoring 72–85 points were placed in the proficient level by some judges *and* in the advanced level by other judges.
- Booklets scoring over 85 points were placed in the advanced level by all judges.

The distribution of scores in each level indicates that the cut score between basic and proficient is somewhere between 46 and 60 points, and the cut score between proficient and advanced is somewhere between 72 and 85 points.

For the pinpointing round, more booklets are selected in the ranges of scores in which the range-finding round placed the cut scores. Typically, 20–30 booklets are selected for each cut score for the pinpointing round. Judges place those booklets into one of the two relevant performance levels. Final recommendations of the cut scores are made following group discussions among the judges. The booklet's scores are hidden from the judges until the end of the session.

Choosing the Cut Score.

The cut score for each performance level is determined by finding the point that best distinguished between two adjacent performance levels as is done for the Contrasting Groups method. As described for that method, smoothing of the distributions of scores may be helpful.

Compromise Method

Both normative and absolute information can be used simultaneously to set reasonable cut scores. Hofstee, for example, provided a method of considering both sources

of information in the early 1980s. Judges determine the lowest cut score that they would consider acceptable even if everybody failed, and the highest cut score that they would still find acceptable even if everybody passed. Then judges select the minimum and maximum acceptable percentages of failures. Using those points and the score distribution for the test, a cut score that fits within the allowable ranges can usually be found. Considering both normative and absolute information in setting a cut score can help avoid the establishment of unreasonably high or low values.

Choosing a Standard-Setting Method *(Much of this section is excerpted from Passing Scores)*

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. The best kinds of data to use 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, use a 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 from sufficiently trained judges. How many students should be judged for a Contrasting Groups method? It is difficult to state a minimum number because it depends on many factors such as the consequences of misclassifying a student, the extent to which judges agree on the classifications of students, and the number of performance levels in which students are classified. The best strategy is to consult a statistician who is familiar with the issues involved in setting cut scores. Though this is not a strict rule, it is a reasonable goal to have at least 100 students classified in each performance level. You may, however, need classifications of many more students in total to have 100 in each performance level. For example, if the performance levels are pass and fail, you may need to classify many more than 200 students to determine the cut score if most students are classified as passing and relatively few are classified as failing.

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, Ebel's or Bookmark, 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, Ebel's, and Bookmark are especially useful when it is important that the cut score represent the standard of large and diverse groups of people. For example, in choosing the cut score on a math test used as a requirement for high school graduation, it may be important to include the opinions of employers, and community leaders. These people are not in a position to observe the mathematical skills of many 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, Ebel's, and the Bookmark method.

Nedelsky's, Angoff's, Ebel's, and the Bookmark 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 cut score 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, Ebel's, and the Bookmark 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 can be a difficult task 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 not dependent on data on the difficulty of each question as is Bookmark. If the required data are available, the Bookmark method is the easiest of the methods to explain and the fastest to use because the judges do not have to state a probability for each question. Angoff's method is next in ease of application. Ebel's method enables 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.

Another consideration should be the types of questions in a test. Nedelsky's method works only for multiple choice questions. It 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. The extended Angoff method and the Bookmark method can be used with tests that contain both multiple-choice and constructed-response questions. The Body of Work method works best when the test consists mostly of constructed-response questions.

Conclusion

It is impossible to prove that a cut score is correct. Therefore, it is crucial to follow a process that is appropriate and defensible. Ultimately, cut scores are based on the opinions of a group of people. The best we can do is choose the people wisely, train them well in an appropriate method, give them relevant data, evaluate the results, and be willing to start over if the expected benefits of using the cut scores are outweighed by the negative consequences.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508–600). Washington, DC: American Council on Education.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, *21*, 147–152.

Cizek, G. J. (1996a). Setting passing scores. *Educational Measurement: Issues and Practice*, *15*(2), 20–31.

Cizek, G. J. (1996b). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, *15*(1), 12–21.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–50.

De Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement*, *22*, 263–269.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hambleton, R. K., & Pitoniak, M. J. (in press). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Washington, DC: American Council on Education.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S.B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109–127). San Francisco: Jossey-Bass.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 485–514). Washington, DC: American Council on Education/Macmillan.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice*, *10*, (2), 3–6, 10, 14.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64* (3), 425–461.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment*, *5* (3), 129–145.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum Associates.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: ETS.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice*, *10* (2), 7–10.

Mitzel, H. C., Lewis, D. M., Patz, R.J., & Green, D.R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Zieky M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51). Mahwah, NJ: Lawrence Erlbaum Associates.

Listening. Learning. Leading.