
RESEARCH METHODOLOGY

Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education

Steven M. Downing

Ara Tekian

Rachel Yudkowsky

Department of Medical Education

University of Illinois at Chicago

Chicago, Illinois, USA

Background: Establishing credible, defensible, and acceptable passing scores for performance-type examinations in real-world settings is a challenge for health professions educators. Our purpose in this article is to provide step-by-step instructions with worked examples for 5 absolute standard-setting methods that can be used to establish acceptable passing scores for performance examinations such as Objective Structured Clinical Examinations or standardized patient encounters.

Summary: All standards reflect the subjective opinions of experts. In this “how-to” article, we demonstrate procedures for systematically capturing these expert opinions using 5 research-based methods (Angoff, Ebel, Hofstee, Borderline Group, and Contrasting Groups). We discuss issues relating to selection of judges, use of performance data, and decision-making processes.

Conclusions: Different standard-setting methods produce different passing scores; there is no “gold standard.” The key to defensible standards lies in the choice of credible judges and in the use of a systematic approach to collecting their judgments. Ultimately, all standards are policy decisions.

Teaching and Learning in Medicine, 18(1), 50–57

Copyright © 2006 by Lawrence Erlbaum Associates, Inc.

Establishing credible, defensible, and acceptable passing or cutoff scores for performance-type examinations in health professions education can be challenging.^{1–3} There is a large literature of standard setting, much of which has been devoted to empirical passing score studies and comparisons of various standard-setting methods that are appropriate for selected-response tests or performance tests used in kindergarten through Grade 12 educational settings.^{4–7} Standards can be categorized as either relative (norm based) or absolute (criterion based). Relative standards identify a group of passing and failing examinees relative to the performance of some well-defined group; the passing score or standard will depend on the performance of the specific group tested—for example, the bottom 5% of the class or those who score more than 1 *SD* below the mean. Ab-

solute standards are based on a predetermined level of competency that does not depend on the performance of the group—for example, a score of 70%. Our purpose in this article is to describe specific procedures to carry out five different absolute standard-setting methods, each of which can be used to establish acceptable passing scores for performance examinations, such as Objective Structured Clinical Examinations (OSCEs) or standardized patient examinations, in health professions settings. This is a “how to” article, with the major emphasis on step-by-step instructions and worked examples of each of the five methods: Angoff,⁸ Ebel,⁹ Hofstee,^{10,11} Borderline Group,⁷ and Contrasting Groups.^{7,12,13} Although each of the standard-setting methods exemplified here is research based, our primary objective in this article is to demonstrate how to apply the methods in

Correspondence may be sent to: Steven M. Downing, Associate Professor of Medical Education, University of Illinois at Chicago, College of Medicine, Department of Medical Education (MC 591), 808 South Wood Street, Chicago, IL 60612–7309, USA. E-mail: sdowning@uic.edu

real-world settings and not to argue the relative merits of one method compared to another method or recommend one method over another method.

We emphasize that the procedures we describe in this article are examples of only one particular way to implement each method. Every setting is unique and minor (or major) modifications to these standard-setting procedures may be required in some settings.

All standard-setting methods require judgment; the very nature of a passing score concerns deciding “how much is enough.” Even normative or relative standard-setting methods require some type of judgment; for example, “At what point on the distribution of scores will the cut point be located?” In essence, a passing score is an operational statement of policy. For the absolute methods we discuss here, the choice of content expert judges is crucial. The passing scores established are only as credible as the judges and the soundness of the systematic methods used.¹ Content expertise is the most important characteristic of judges selected for the standard-setting exercise. Judges must also know the target population well; understand both their task as judges and the content materials used in the performance assessment; be fair, open-minded, and willing to follow directions; be as unbiased as possible; and be willing and able to devote their full attention to the task. In some settings, it may be important to balance the panel of judges with respect to demographic variables such as ethnicity, gender, geography, and subspecialization. For most methods and settings, 5 to 6 independent judges might be considered minimum, with 10 to 12 judges the maximum. Practical considerations must often play a major role in judge selection, the numbers of judges used, the venues for standard-setting exercises, and the exact manner in which the procedures are implemented.

Most absolute standard-setting methods pivot on the idea of the borderline student or examinee. This concept originated with Angoff’s⁸ original work on absolute passing scores. The cut score separating those who pass from those who fail corresponds to the point that exactly separates those who know (or can do) just enough to pass from those who do not know enough (or can not do enough) to pass. The borderline examinee is thus one who has an exactly 50:50 probability of passing or failing the test. The borderline examinee is the marginal student—one who on some days might just barely pass your assessment but on other days might fail.

The definition of borderline examinee is straightforward, but operationalizing this definition can be challenging. Asking judges to describe borderline students they have known imparts a clear understanding of what it means to be borderline and facilitates group consensus prior to beginning the standard-setting work.

Should Absolute Standard-Setting Methods Be Completely “Absolute?”

The term *absolute* implies that passing score judgments are made such that judges are blind to all actual performance data, looking only at the content of the performance assessment or the stimuli or prompts. This is the purist view, but the reality is that totally pure absolute standards rarely turn out to be realistic, acceptable, or useful in the real world of health professions education. Expert judges tend to expect even borderline examinees to know more and be able to do more than is realistic. Many studies⁴ have demonstrated that judges, absent all performance data, tend to set unrealistically high passing scores, which will fail an unreasonably high proportion of students. Experts almost always expect too much of novice learners.

The point of view we adopt in this article is that judges must be “calibrated” to have a realistic expectation of actual student performance. Such calibration requires presenting some performance data to judges so that standard-setting panels have a reasonable expectation concerning actual student performance on the assessment. Many experts in education disagree with this point of view and may label such methods biased. We prefer the efficiency of judge calibration to the inefficiency of repeating the standard-setting exercise a second or third time if the first rounds result in unacceptably high standards.

Systematic Methods

There is no “gold standard” for a passing score. There is no perfect passing score “out there” waiting to be discovered. Rather, the passing score is whatever a group of content expert judges determine it is, having followed a systematic, reproducible, absolute, and unbiased process. The key to defensible and acceptable standards is the implementation of a careful, systematic method to collect expert judgments, preferably a method that is based on research evidence. Different standard-setting methods will produce different passing scores; and different groups of judges, following exactly the same procedures, may also produce different passing scores for the same assessment. Such facts are troubling only if one expects to discover the perfect or gold standard passing score. Process is the key concept, remembering that all passing scores are ultimately policy decisions, which are inherently subjective.¹⁴

We discuss five different methods here—Angoff,⁸ Ebel,⁹ Hofstee,^{10,11} Borderline,⁷ and Contrasting Groups^{7,12,13}—all of which are potentially useful for establishing realistic and acceptable standards for performance examinations in the health professions.

Some Tips

It is essential that every standard-setting judge fully understand the relation between passing scores and passing rates. The *passing score* is the score needed to pass the performance test, often expressed as a percent-correct score. The *passing rate* is the percentage of students who pass the test at any given passing score (sometimes expressed as the failure rate). The higher the passing score, the lower the passing rate. If standard-setting judges confuse these two statistics, their judgments will confuse the passing score and become a threat to the validity of the standard.

In a performance exam that includes several components or stations, such as a standardized patient encounter or OSCE, the question of compensatory versus noncompensatory standards must be addressed. Can good performance on one station, or one component, compensate for poor performance on another? If so, the overall standard will consist of the simple average of standards across stations or components (compensatory scoring). In some cases, however, a noncompensatory approach may be more appropriate to ensure that students reach a minimum level of competence in several crucial but different domains—such as clinical reasoning and communication skills. In this case, standards must be set separately for each domain. In clinical cases, faculty often feel very strongly that a few crucial items must be accomplished for the student to pass, regardless of overall score. These items should be discussed at both the scoring and standard-setting stages of exam planning. Such items must comprise a sufficiently large sample of student behavior to be reliable because very small samples of items—containing large sampling error—may result in incorrect decisions.

Some panels of judges want lots of performance data to help them calibrate their judgments, whereas others do not wish to have such data or minimize such data. Either condition is fine as long as there is some consistency in the conditions. Some judge panels wish to know, from time to time throughout the process, what passing score and/or passing rate they have established thus far in the process. Again, this is a matter for professional judgment, and we take the position that, in general, more data is better than less data for all judgments, realizing that this is a minority opinion in the educational world but acknowledging the fact that all passing scores are ultimately policy decisions.

The Angoff⁸ Method

The Angoff⁸ method of establishing absolute passing standards has a well-established history of research and is easily adapted to performance examination prompts.⁶ In this method, content experts make judgments about every prompt or checklist item, so it is

fairly easy to defend the resulting passing scores. The Angoff⁸ method was the first of the absolute methods and thus has the longest history of successful use, even in high-stakes testing situations.

Angoff⁸ Standard-Setting Procedures

There are five steps in implementing an Angoff⁸ standard-setting exercise:

1. The standard-setting judges discuss the characteristics of a borderline examinee and note specific examples of borderline students.
2. Judges come to a consensus agreement on the qualities of the borderline examinee, with specific examples in mind.
3. Each judge estimates the performance of the borderline examinee for each performance prompt, item, or rating (0% to 100%).
4. These judgments are recorded (usually by a nonjudge recorder or secretary).
5. Judgments are then systematically combined (totaled and averaged) to determine a passing score on the performance test.

Some actual performance data may be given to the judges. Summary data such as the mean and standard deviation of the standardized patient case or OSCE will help to calibrate judges as to the difficulty of the case for real students. Alternately, more specific data may also be presented such as the proportion of the total group of students who get a checklist item correct.

Item Review and Rating

Judgments are carried out at the prompt level (checklist item or rating-scale item) for each case. Passing scores are thus computed for each case; averaging over the case passing scores determines the overall passing score for all cases in the performance examination.

The prompt or checklist item review begins with one of the judges reading the first item on the checklist. First the reader and then the other judges on the panel give their estimate of how well a borderline candidate will score on that item; judges rotate clockwise for each new item. Each judge's estimate (judgment) is recorded on a recording sheet or a computer spreadsheet. For each checklist item/prompt, the judges answer one of the following two equivalent questions:

1. How many individuals in a group of 100 borderline examinees will perform this checklist item correctly (0% to 100%)? Or
2. What is the probability that one borderline examinee will perform this checklist item correctly (0 to 1.0)?

Note that the Angoff⁸ question asks judges to estimate how well students will perform, not how well they should perform. The difference between will and should needs to be emphasized. If the judgments for an item differ by 20% or more, those judges who provided the high and low scores may lead a discussion of their ratings for that item. Throughout the process, judges can modify their ratings or judgments. The review and rating of prompts continues until the entire checklist has been completed.

Table 1 shows the Angoff⁸ ratings for a 10-item performance examination rated by seven Angoff⁸ judges. The case passing score (percent) is the simple average of passing scores for all items.

A variant of the Angoff⁸ method (actually Angoff's⁸ original method) is to ask judges to make a simple "yes" or "no" judgment about each item/prompt. The question becomes, "Will the borderline examinee respond correctly to this item?" All "yes" answers are coded as 1, with "no" answers coded 0. The simple sum of the 1s and 0s becomes the raw passing score when averaged over all judges (See Table 2). This simplified Angoff⁸ method (direct or yes-no method) may be useful for some types of examinations, such as laboratory tests, for which use of the traditional Angoff⁸ method would be difficult.¹⁵⁻¹⁷

Another variant of the Angoff⁸ method for use with a rating scale rather than a dichotomous checklist would be to have each judge independently estimate the rating that a borderline student will get on each item. For example, if the student is being rated on a 5-point scale ranging from 1 (*poor*) to 5 (*excellent*), a borderline student might be expected to achieve a rating of 3 on Item 1 and a 2 on Item 2. Calculate the mean rating for each item across all judges and average over items to obtain the raw passing rating score.

Table 1. Sample Angoff Ratings and Calculation of Angoff Passing Score

Item	Rater							M
	1	2	3	4	5	6	7	
1	.80	.87	.85	.90	.80	.95	.85	0.86
2	.70	.75	.80	.85	.75	.85	.75	0.78
3	.50	.63	.55	.60	.65	.60	.60	0.59
4	.70	.68	.70	.70	.65	.70	.70	0.69
5	.75	.70	.80	.85	.70	.85	.80	0.78
6	.60	.65	.80	.75	.65	.85	.80	0.73
7	.50	.58	.55	.60	.70	.90	.60	0.63
8	.70	.78	.75	.75	.65	.80	.70	0.73
9	.45	.50	.50	.45	.43	.55	.45	0.48
10	.60	.69	.65	.65	.65	.70	.70	0.66
Sum ^a								6.93
Pass Score ^b								69.30%

^aRaw Passing Score = Sum of item means = 6.93. ^bPercent Passing Score = 100% × (sum of item means/number of items) = 100% × (6.93/10) = 69.30%.

Table 2. Sample for Simplified/Direct Angoff Ratings and Calculation of Passing Score

Item	Rater					M
	1	2	3	4	5	
1	1	1	0	1	1	0.8
2	1	1	1	1	1	1.0
3	1	0	1	0	1	0.6
4	0	0	0	0	0	0.0
5	0	0	0	1	1	0.4
6	1	1	1	1	1	1.0
7	0	0	1	0	1	0.4
8	1	1	1	1	1	1.0
9	1	1	1	1	0	0.8
10	0	0	1	0	0	0.2
Sum ^a						6.2
Pass Score ^b						62%

^aRaw Passing Score = Sum of 0/1 means = 6.2. ^bPassing Score Percent = 100% × (sum of item means/number of items) = 100% × (6.2/10) = 62%.

The Ebel⁹ Method

The Ebel⁹ method requires judges to consider both the difficulty of the item and its relevance. This method gives standard-setting judges more information about the performance test and its individual items but also requires more work and time of the judges than some other methods.

Ebel⁹ Standard-Setting Procedures

There are two major tasks required to implement an Ebel⁹ standard-setting procedure:

1. Prepare a matrix of item numbers categorized by relevance and difficulty.
2. Estimate the proportion of borderline examinees who will succeed on the type of item in each cell in this matrix.

Item difficulty is determined by calculating the average difficulty (percent correct) for each item based on actual data from an administration of the performance exam to a (representative) group of examinees. Difficulty ranges (easy, medium, hard) are arbitrarily determined but should have some rational basis in the empirical data.

Relevance ratings (essential, important, acceptable) for each item must be obtained from judges (see Number 6 in the following list). It is customary for the same judges used to give the final Ebel⁹ ratings to carry out the relevance ratings, but this is not essential. Also, because some time is needed to carry out various computations and to create rating forms once the relevance ratings are obtained, it may be necessary to divide the Ebel⁹ standard-setting exercise into two separate ses-

sions. A different group of judges could carry out relevance ratings if circumstances warrant.

Here is a summary of steps to accomplish an Ebel⁹ standard-setting exercise:

1. Familiarize the judges with the content of the performance cases and the checklists or rating scales.
2. Discuss specific definitions of the relevance categories used: “essential,” “important,” and “acceptable.” For example, “essential to good patient care—if this item is not accomplished, the patient’s health is at risk.”
3. Have each judge rate each item as essential, important, or acceptable.
4. Compute summary statistics (average across judges) for the relevance ratings of each item.
5. Compute mean item difficulty (proportion correct) for each item or prompt of each case or station based on actual performance data.
6. For each case, prepare a matrix of items sorted by relevance and difficulty (see Table 3).
7. Lead the judges in a discussion of borderline student performance.
8. Reach some common understanding of the characteristics of the borderline examinee.
9. Ask each judge to provide an answer to the following question for each set of items designated by a cell in the matrix: “If a borderline student had to perform a large number of items or prompts like these, what percentage (0% to 100%) would the student perform correctly?”
10. Each judge records the estimated percentage of students who will correctly perform items like those noted in the cell.
11. Average judgments across all judges are computed and recorded as shown in Table 3.
12. A weighted mean, defined as the number of items in the cell multiplied by the mean rating for that cell, is computed for each row of the matrix and then summed.
13. Adding the total for each row of the matrix gives the raw passing score as determined by the Ebel⁹ judges.

The Hofstee^{10,11} Method

The Hofstee^{10,11} method is sometimes referred to as the “relative-absolute compromise method” because it combines features of both relative and absolute standard setting.^{10,11} Judges are asked to define minimum and maximum acceptable passing scores and failure rates. The standard is determined by the midpoint of the cumulative frequency curve of the exam scores as it passes through this bracketing rectangle (See Figure 1). Like the Ebel⁹ method, the Hofstee^{10,11} method requires analyzing and summarizing performance data from the test prior to collecting judgments. Alternatively, performance data can be obtained from a subgroup of representative examinees or from a prior administration of the examination. If the judges do not take actual performance data under close consideration, the cumulative frequency distribution curve may not be included within the score boundaries they define. The Graphical Hofstee^{10,11} (see During the Exercise, procedure Step 6, alternate) avoids this problem and ensures that the standard-setting exercise will result in judgments that are applicable to the specific group examined.

Some researchers discourage use of the Hofstee^{10,11} method for high-stakes examinations, perhaps feeling that it is less credible because the judgments are global rather than based on individual items.⁶

Hofstee^{10,11} Standard-Setting Procedures

A group of content-expert judges who are familiar with the students and the performance examinations under consideration are assembled and trained in the Hofstee^{10,11} method.

Before the Exercise

1. Based on actual performance data, compute the mean and standard deviation of the test and any other statistics (such as mean scores at quartile cutoffs) that would be helpful in describing the overall performance of students on the test.
2. Consider presenting graphical data showing the overall distribution of scores.

Table 3. Sample Ebel Ratings and Calculation of Passing Score Matrix of Checklist Item Relevance by Difficulty

Item Relevance	Easy (.80–.99)		Medium (.45–.79)		Hard (0–.44)		Weighted Mean
	Item Number	% Correct	Item Number	% Correct	Item Number	% Correct	
Essential	4, 5	93 ^a	1	81	3	63	2 (.93) + .81 + .63 = 3.30
Important	2	89	10	76	9	59	.89 + .76 + .59 = 2.24
Acceptable	N/A	N/A	7	62	6, 8	42	.62 + (2 (.42)) = 1.46

Note: Raw passing score = sum of weighted m = 3.30 + 2.24 + 1.46 = 7.0 raw points. Percent passing score = 100% × (sum of item means/number of items) = 100% × (7.0/10) = 70 %.

^aIn this example, for two items rated as essential and easy, 93% correct represents the mean judgment of all the Ebel judges.

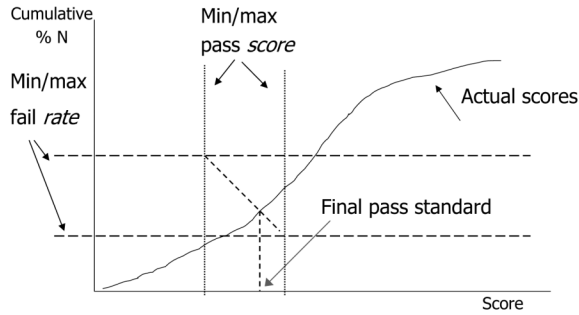


Figure 1. The Hofstee method.

3. Optionally, calculate and present other examination data such as any historical data about student performance on the same or similar tests over time.
4. Calculate and graph the cumulative frequency distribution (as a cumulative percent) of the total performance test score for each case. (Statistical software such as SPSS can be used to plot the cumulative frequency percent.) See Figure 1 for an example.

During the Exercise

1. Present and discuss the data discussed previously with the standard-setting judges.
2. Review the cases and the items, the scoring methods, and other relevant details of the exam.
3. Discuss the borderline examinee with the group of judges, coming to a consensus agreement on the characteristics of the examinee who just barely passes or just barely fails.
4. Present and discuss the four Hofstee^{10,11} questions, ensuring that each judge fully understands each question (see Number 6 following) and its implications.
5. Consider doing a practice run to be certain that judges fully understand the Hofstee^{10,11} procedures.
6. Have each judge answer each of the four questions, as noted here:
 - a. The **LOWEST** acceptable percentage of students to FAIL the examination is: _____ percent (minimum fail rate).
 - b. The **HIGHEST** acceptable percentage of students to FAIL the examination is: _____ percent (maximum fail rate).
 - c. The **LOWEST** acceptable percent-correct score that allows a borderline student to pass the examination is: _____ percent (minimum passing score).
 - d. The **HIGHEST** acceptable percent-correct score required for a borderline student to pass the examination is: _____ percent (maximum passing score).

6. (alternate) Alternatively, have judges draw lines designating the highest and lowest acceptable pass scores and fail rates directly on the cumulative score graph, with instructions to be sure to include the cumulative score line within the rectangle thus defined (Graphical Hofstee^{10,11}). Have judges specify and record the exact numerical value represented by their lines.

After the Exercise

1. Compute the mean percentage for each of the four questions across all judges.
2. Plot the mean of the four data points (minimum and maximum acceptable fail rate and pass score) on the cumulative frequency distribution.
3. The midpoint of the intersection of the minimum and maximum fail rates and pass scores represents the overall passing score for the group of judges. See Table 4 and Figure 2 for a worked example.

If the cumulative frequency distribution curve does not fall within the score boundaries defined by the judges, and the judges cannot be recalled to run the exercise again, the standard can default to the minimum acceptable passing score or the maximum acceptable failure rate determined by the judges. Use of the Graphical Hofstee^{10,11} method (Step 6[alternate] in the preceding list) will help prevent this problem because judges can immediately see the results of their judg-

Table 4. Sample Hofstee Ratings and Calculation of Passing Score

	Rater				M
	1	2	3	4	
Minimum Passing Score	65	70	60	60	64
Maximum Passing Score	75	75	65	70	71
Minimum Fail Rate	5	0	10	7	6
Maximum Fail Rate	20	25	30	30	26

Note: Use rater means to obtain pass score by graphing onto cumulative percent graph (see Figure 2).

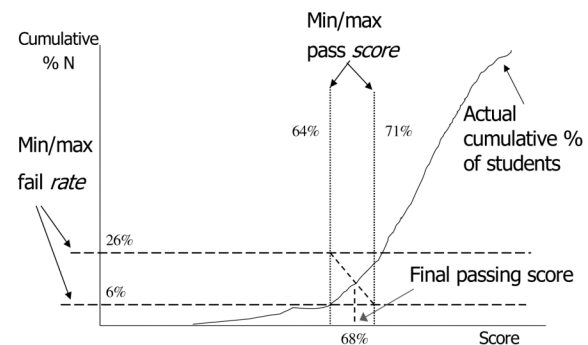


Figure 2. Hofstee example.

ments and whether the cumulative score line falls within the defined boundaries.

Borderline Group Method⁷

The Borderline Group method⁷ is an examinee-centered rather than an item-centered method: Judgments are made about individual test takers, not test items or content. The method can be used only when content experts who are qualified to serve as standard setters (e.g., faculty) directly observe the performance test. (Appropriately trained standardized patients may be considered content experts in communication and interpersonal skills). The observing judges' global ratings are used to determine the checklist score that will be used as the passing standard.

Borderline Group⁷ Standard-Setting Procedures

1. Prepare judges by orienting them to the station or case and to the checklist or other rating instruments.
2. Judges directly observe the test performance of each examinee. Each judge should observe multiple examinees on the same station rather than following an examinee across several stations. The test performance observed may, with appropriate training, consist of performance products such as individual checklist item scores or postencounter notes.
3. The observing judge provides a global rating of the overall performance of each examinee on a 3-point scale: 1 = *Fail*, 2 = *Borderline*, 3 = *Pass*.
4. The performance is also scored (by the judge or another rater) using a multiple-item checklist or rating scale.
5. The mean checklist score of those examinees rated borderline becomes the passing score for the test. (See Figure 3.)

Contrasting Groups Method^{7,12,13}

The Contrasting Groups method^{7,12,13} is another examinee-centered standard-setting method that requires using an external criterion or other method to di-

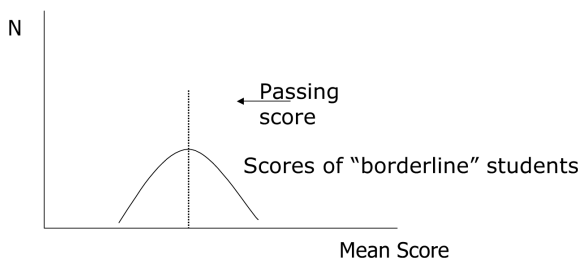


Figure 3. Borderline group method.

vide examinees into two groups: experts versus novices and passers versus failers or competent versus noncompetent. The standard is the score that best discriminates between the two groups. One of the advantages of this method is that the standard can easily be adjusted to minimize errors in either direction. Thus, if the error of greatest concern is mistakenly categorizing an examinee as a “pass” when they should have failed (e.g., in certifying examinations), the standard can be moved to the right. (See Figures 4 and 5.)

Contrasting Groups^{7,12,13} Standard-Setting Procedures

1. Examinee performance is scored by judges or other raters using a multiple-item checklist or rating scale.
2. Examinees are divided into expert and nonexpert groups based on an external criterion or by having expert observers provide a global Pass–Fail rating of the student's overall performance.
3. Graph the checklist score distributions of the two groups.
4. The passing score is set at the intersection of the two distributions if false-positive and false-nega-

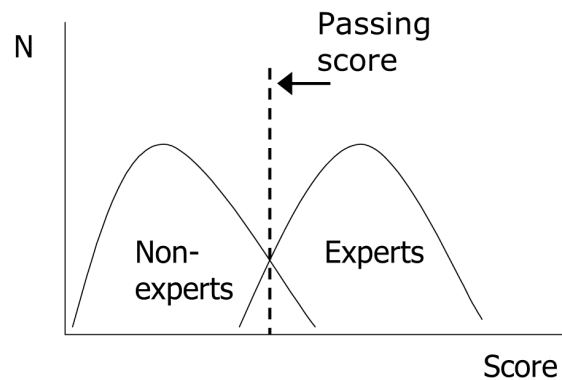


Figure 4. Contrasting groups.

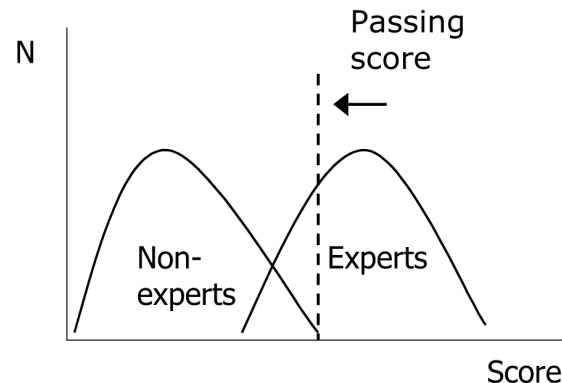


Figure 5. Minimizing passing errors.

Table 5: Comparison of Five Standard Setting Methods

	Judgment Focused on:	Judgments Require Prior Exam Data?	Requires Expert Observers of Performance?	Timing of Judgments
Angoff	Test items	No	No	Before exam
Ebel	Test items	Yes	No	After exam
Hofstee	Whole test	Yes	No	After exam
Borderline Group	Examinee performance	No	Yes	During exam
Contrasting Groups	Examinee performance	No	Yes	During exam

tive errors are of equal weight, or moved to the right or the left to minimize the error of greater concern.

Conclusion

In this article, we have described the procedures for five different methods of setting standards. Which method should you choose for your examination? Frequently, the choice will depend on the practical realities of the test. If content experts such as faculty observe and rate the performance of the examinees, you can choose to use the Borderline Group⁷ or Contrasting Groups^{7,12,13} method. These methods are convenient and simple to implement; faculty are very comfortable with making judgments about an individual performance, and all judgments are made in the course of the exam, so no additional faculty time is needed. If you do not have expert raters observing the exam, only methods involving judgments about the test items or test content (Angoff,⁸ Ebel,⁹ or Hofstee^{10,11}) can be used. See Table 5 for a comparison of the five methods across several important dimensions.

No matter which standard-setting method you choose, some evaluation of the resulting standard is appropriate. Is your cut score acceptable to your stakeholders? If not, is it because the test was not appropriately constructed, because your curriculum did not prepare students for the exam, or because your standard-setting judges did not have (or use) information about the actual performance of the students? Formal approaches to assessing the psychometric characteristics of standards have been proposed¹ but are beyond the scope of this article.

Different standard-setting methods will produce different passing scores; there is no gold standard. The key to defensible standards lies in the choice of credible judges and in the use of a systematic approach to collecting their judgments. Ultimately, all standards are policy decisions reflecting the collective, subjective opinions of experts in the field.

References

1. Norcini JJ, Shea JA. The credibility and comparability of standards. *Applied Measurement in Education* 1997;10:39–59.

2. Friedman M. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher* 2000;22:120–30.
3. Norcini J, Guille R. Combining tests and setting standards. In GR Norman, CPM Van der Vleuten, DI Newble (Eds.), *International handbook of research in medical education* (pp. 811–34). London: Kluwer Academic, 2002.
4. Cizek GJ. *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2001.
5. Cizek GJ, Bunch MB, Koons H. Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice* 2004;23(4):31–50.
6. Norcini JJ. Setting standards on educational tests. *Medical Education* 2003;37:464–9.
7. Livingston SA, Zieky MJ. *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service, 1982.
8. Angoff WH. Scales, norms, and equivalent scores. In RL Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education, 1971.
9. Ebel RL. *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall, 1972.
10. Hofstee WKB. The case for compromise in educational selection and grading. In SB Anderson, JS Helmick (Eds.), *On educational testing* (pp. 107–27). San Francisco: Jossey-Bass, 1983.
11. De Gruijter DN. Compromise models for establishing examination standards. *Journal of Educational Measurement* 1985;22:263–9.
12. Burrows PJ, Bingham L, Brailovsky CA. A modified contrasting groups method used for setting the passmark in a small scale standardized patient examination. *Advances in Health Sciences Education* 1999;4:145–54.
13. Clauser BE, Clyman SG. A contrasting-groups approach to standard setting for performance assessments of clinical skills. *Academic Medicine* 1994;69:S42–S44.
14. Kane MT, Crooks TJ, Cohen AS. Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education* 1999;4:195–207.
15. Downing SM, Lieska NG, Raible MD. Establishing passing standards for classroom achievement tests in medical education: A comparative study of four methods. *Academic Medicine* 2003;78:S85–87.
16. Subhiyah RG, Featherman CM, Hawley JL. *How to set pass/fail standards on examinations*. Workshop presented at the Annual Meeting of the Generalists in Medical Education, San Francisco, November 2002.
17. Impara JC, Plake BS. Standard setting: An alternative approach. *Journal of Educational Measurement* 1997;34:353–66.

Received 20 June 2005

Final revision received 18 July 2005