10. Establishing Passing Standards

Craig N. Mills
Educational Testing Service

Follow this and additional works at: http://digitalcommons.unl.edu/buroslicensure

Part of the Adult and Continuing Education and Teaching Commons, Educational Assessment, Evaluation, and Research Commons, and the Other Education Commons


This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Licensure Testing: Purposes, Procedures, and Practices by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
INTRODUCTION

When tests are used to determine eligibility for a license, a passing standard or cut score must be established that divides the test scores into two categories: eligible for license or not. Standard setting has been widely researched and there are many reviews available (see, for example, Jaeger, 1989; Mills & Melican, 1988; Berk, 1986; Hambleton, 1980; Hambleton & Eignor, 1980; and Shepard, 1980a, 1980b), yet there is limited practical advice available for conducting standard setting studies and establishing standards. The one available resource (Livingston & Zieky, 1982) is somewhat dated. The purpose of this chapter is to provide a practical discussion of the entire standard setting process. The steps in a standard setting study are explained. Commonly used standard setting methods are described, examples are provided, and the methods are critiqued. Procedures for conducting a standard setting study and adjusting the resulting preliminary standard are also explained. The chapter also discusses factors other than test performance that can be considered in setting standards on licensure tests.

1The author wishes to express his appreciation to Jay Breyer, Jim Impara, Skip Livingston, Jerry Melican, Maria Potenza, Nancy Thomas-Ahlwahla, and Michael Zieky who, despite their disagreement with some of my positions, provided valuable reviews of this chapter.

2This chapter assumes that other important steps in the test development progress (e.g., establishing test specifications, conducting a job analysis) have already been completed. These steps are discussed in other chapters.
Standard setting is a multiple-step process involving different groups. There are typically three groups involved in the process: the “test sponsor,” the investigator, and expert raters (or judges). The term “test sponsor” refers to the organization (e.g., licensure board), that has ultimate responsibility for the testing program. Although the sponsor may contract for testing services (test development, administration, and statistical analysis), it bears responsibility for the soundness of the test and testing program and has policy, financial, and legal responsibility as well. The investigator is the individual (or group) responsible for conducting the standard setting study and advising the test sponsor on all aspects of it. The investigator may be an employee of the sponsor, a testing services provider, or an independent consultant. The investigator’s responsibilities extend from initial discussion of the design of the study through the actual data collection and analysis, and extend (typically) to acting as a resource during the deliberations leading to the establishment of the operational standard. Expert raters are typically educators and/or practitioners in the field who are convened on one or more occasions to provide judgments about the test, examinees, and (possibly) the appropriateness of the recommended standard.

It is important to identify clearly which parties are involved in each step and what their specific responsibilities are. For example, test sponsors will often use an external investigator to conduct the standard setting study. This is sound practice if standard setting expertise is not available within the sponsor’s organization, but does not exempt the sponsor from the responsibility of establishing the final standard. Figure 1 lists the steps in establishing a standard and the parties involved in each step. Each step is explained in the remainder of the chapter.

Determine the Need for a Standard

In most licensure settings, the decision to develop a test is based on the need to make decisions about individuals (e.g., the individual has sufficient knowledge and skills to receive a license or not). However, it is important that the development of the licensure test itself is justified. It is appropriate, for example, for a legislative body to decide that there is sufficient risk to the public from ill-prepared practitioners that a test to distinguish between individuals who can provide appropriate service and those who cannot is necessary.

Livingston and Zieky (1982) suggest that test sponsors be prepared to justify the use of a standard. Although it may be true that fairer licensure decisions will result from the program than from a case-by-case consideration of applications, it is likely that there will be resistance to the imposition of a test. Test sponsors should know the likely criticisms and be ready to respond to them and contrast the fairness of the program with current practice. Several other issues should be considered as well. The sponsor should ensure that the appropriate reliability and validity analyses will be conducted. Administrative procedures should be addressed. For example, how often will individuals be allowed to test? Will periodic license renewal be required? Will current practitioners be “grandfathered” into the program? Under what conditions (if any) should exceptions be granted? How much advance notice will be given of the requirement to pass the test? These issues
Determine the Need for a Standard
Design the Study
   Selecting a Standard Setting Method
      Normative Standards
      Absolute Standards
         Arbitrary Standards
         Absolute Methods: Evaluation of the Test
         Absolute Methods: Evaluation of Individuals
      Setting Standards on Performance Assessments
         Simple performance assessments
         Complex performance assessments
Planning Study Procedures and Analyses
   Multiple Iterations
   Providing Feedback on the Ratings
   Discussion of Ratings
   Placing Limits on the Judgments
   Adjusting Ratings for Guessing
   Providing Feedback on Examinee Performance
   Timing of the Ratings
   Item Criticisms
Select Expert Raters
Conduct the Study
   Introductory Session
   Defining Minimal Competence
   Training the Raters
Evaluate Results and Establish Standard
   Adjusting the Standard
      Standard Error of Measurement Adjustment
      Observed Score Distribution Adjustments
      Other Factors That May be Considered
Document the Study

Figure 1. Steps in standard setting

are more directly related to the operational aspects of the testing program than to the establishment of a standard, but can affect standards. Interested readers are referred to Livingston and Zieky (1982) for a discussion of these and other issues.

DESIGN THE STANDARD SETTING STUDY

As is true in any inquiry, the design of the standard setting study is critical. Important considerations include selecting a standard setting method, identifying the data collection methodology, specifying analyses, and ensuring that the expert judges will have appropriate information and training, and the individuals representing the sponsor (i.e., the board) are aware of their responsibility in setting the standard.
SELECTING A STANDARD SETTING METHOD

Standard setting methods fall into two broad categories, normative and absolute. *Normative standards* limit the number of individuals eligible for licensure by specifying a *percentage or number* of individuals who will be granted a license or by specifying a point in the distribution of scores as the standard (e.g., one standard deviation below the mean or the 55th percentile of the national norm group). *Absolute standards* are set to specify a specific required *level of performance* on the test. All individuals who attain that level of performance are granted a license, regardless of the number or percentage of individuals falling above or below the standard.

**Normative Standards**

An advantage of normative methods is that the passing rate is known before the test is administered. This can be useful when, for example, financial awards are based on test results and only a limited number of awards are available. For example, a scholarship or fellowship program might have a fixed amount of money to award and a set amount for each award. Awards will be granted to the individuals with the highest test scores until the funds are exhausted. Consider, for example, the test scores shown in Table 1. Suppose a university scholarship program has sufficient funds to support the six “most deserving” new students based solely on test scores (not a recommended practice, but used here for illustrative purposes). In the first year, awards are made to students receiving test scores of 93 and above, however, in the second year, the cut off is 96. If the rewards available are limited, it can be appropriate to use normative methods. These methods can also be used in a two-step selection process. For example, a test might be used to select some fixed number of individuals within the examinee group who would then proceed through an extensive interview process as finalists in a multi step assessment program.

<table>
<thead>
<tr>
<th>Student</th>
<th>Year 1</th>
<th>Year 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>2</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>3</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>4</td>
<td>96</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>94</td>
<td>96</td>
</tr>
<tr>
<td>6</td>
<td>93</td>
<td>96</td>
</tr>
<tr>
<td>7</td>
<td>92</td>
<td>95</td>
</tr>
<tr>
<td>8</td>
<td>90</td>
<td>93</td>
</tr>
<tr>
<td>9</td>
<td>87</td>
<td>90</td>
</tr>
<tr>
<td>10</td>
<td>85</td>
<td>86</td>
</tr>
</tbody>
</table>

Table 1. Scores of the Top 10 Examinees in 2 Years
10. ESTABLISHING PASSING STANDARDS

In most licensure situations, however, the intent is not to select a limited number of individuals, but rather it is to verify whether or not each individual should receive the benefits accorded to those who demonstrate at least “minimal competence.” There is, therefore, typically no reason to limit the number of individuals passing the test. In fact, use of a normative procedure will not guarantee that all individuals who pass the test have similar levels of skill. If a test is administered to a particularly able group of examinees, some able individuals will not pass simply because there are so many high scoring examinees. Conversely, if the examinee group is not particularly able, some with relatively low scores will pass. Suppose the test results in Table 1 were for a licensure exam. The seventh highest scoring examinee in Year 2 seems deserving of licensure if the sixth person in Year 1 is. For this reason, normative standards are typically inappropriate in licensure settings.

Absolute Standards

Absolute standards are used to make judgments about each individual’s test performance without regard to other individuals who have taken the test. Returning to Table 1, for example, suppose the standard was set at 96. In Year 1, only three individuals would pass the test. In Year 2, however, six examinees would pass. Regardless of the ability of the group tested, individuals demonstrating “acceptable” performance would be licensed each year.

Absolute standard setting methods fall into three broad categories: arbitrary methods, methods based on evaluation of test content, and methods based on judgments of expected or observed examinee performance.

Arbitrary Standards

Arbitrary standards are established without regard to test content and difficulty. A test sponsor might, for example, make a statement such as “70% represents passing in most courses, so 70% will be the cut off on the test.” Arbitrary standards have, appropriately, fallen into disuse. The primary reason these standards are inappropriate is that they do not take into account any characteristics of the test-taking population, the test, or the interaction between the two. As a result, the standards are likely to be unfair to some or all test takers.

Absolute Methods Based on Evaluation of the Test

The most commonly discussed standard setting methods based on evaluation of test content are the Nedelsky (1954), Angoff (1971), Jaeger (1978), and Ebel (1979) methods. These methods all require subject matter experts to rate every item in the test. With the exception of the Jaeger method, the methods also require estimation of the difficulty of items (or sets of items in the Ebel method) for a hypothetical group of “minimally competent” examinees.

---

3The term “arbitrary standard” is used in a specific sense here. All standard setting decisions are arbitrary in some sense. This does not, however, necessarily imply capriciousness. An arbitrary decision can be based on consideration of many factors associated with the test and the conditions under which it is being used. In this section, arbitrary means that the standard is set without regard to any of these factors.
The Nedelsky Method. Raters using the Nedelsky method evaluate each answer option of a multiple-choice question to predict whether or not the “minimally competent examinee” would identify it as incorrect. Item difficulty for those examinees is then estimated by assuming that they guess randomly among the remaining options. Because the rating task requires evaluation of the attractiveness of each option, the Nedelsky method ensures consideration of each component of each item (the question, incorrect options, and correct answer).

A modification of the Nedelsky procedure allows judges to rate distractors as “uncertain” (Saunders, Ryan, & Huynh, 1981). In this case, it is assumed that the minimally competent examinee will eliminate these distractors half of the time. The probability that the minimally competent examinee will provide a correct response is calculated similarly to the more common method, but “uncertain” distractors count as half an option.

An example of the Nedelsky method as it is typically implemented is depicted in Figure 2. The figure shows one rater’s evaluation of 10 multiple-choice questions. The first five items are five-option items and the remainder contain four options. For Item 1, the rater eliminated options A, C, and D, predicting that the minimally competent examinee would be able to identify those options as clearly incorrect. Thus, predicted item difficulty is .50 (assuming that minimally competent examinees guess randomly between the two remaining options). Probabilities are determined similarly for all items and summed to determine the expected test score of the minimally competent examinee. The average of these scores across raters is the initial estimate of the cut score.

There are at least four drawbacks to the method. First, it can only be used with multiple-choice tests because each distractor must be rated. Second, the assumption that examinees eliminate clearly incorrect options and then guess randomly among the remaining options does not reflect typical test taking behavior (Melican, Mills, & Plake, 1987). Third, some types of items (e.g. “multiple multiple choice” items) are difficult to rate (Melican & Thomas, 1984). Finally, the estimated item difficulties cannot vary along the full range of difficulty, but are limited to discrete points on a non-symmetrical scale (Brennan & Lockwood, 1980). For a four-option multiple-choice question, for example, the only possible estimates of item difficulty are .25, .33, .50, and 1.00. Despite these drawbacks, the Nedelsky method remains popular in certain professions (although its popularity appears to have declined in recent years).

The Angoff Method. Raters using the Angoff method estimate the difficulty of each item for a hypothetical group of minimally competent examinees, usually by estimating the proportion of such a group that would answer the item correctly. The estimated cut off score for a judge is calculated by summing the item difficulty estimates.

---

4The Nedelsky method, as first published, required consensus among the raters on each distractor.

5Multiple multiple choice items typically present a list of possible answers of which one or more may be correct. Examinees must first identify which answers are correct and then locate the option that contains all correct answers.
An example of the Angoff method is shown in Figure 3. Ratings of five experts for 10 items are shown. Cut scores range from 5.20 to 7.25 and average 6.59. Thus, the estimated cut score is seven items correct.

The Angoff method is the most commonly used standard setting method (Sireci & Biskin, 1992). Ratings are easily obtained, calculation of a cut score is simple, and the method can be easily explained. However, the method also has drawbacks. Raters may judge item difficulty solely on the stem of the item. Because distractors play an important role in item difficulty, raters who do not evaluate them carefully may over- or underestimate item difficulty. Furthermore, even with extensive training, the correlation between raters' estimates of item difficulty and actual item difficulty are often low (Melican & Mills, 1987; Cross, Impara, Frary, & Jaeger, 1984). Other criticisms include the subjectivity of the ratings, concern with the reliability of the method, and the sensitivity of the method to the level of expertise of the judges (Maurer, Alexander, Callahan, Bailey, & Dambrot, 1991).

There are several variations of the Angoff method. Commonly, data collection is simplified by providing raters with a fixed number of equally spaced data points to estimate performance of the minimally competent group (Bernknopf, Curry, & Bashaw, 1979). Some variations limit the number of estimates available, but use a non-symmetric scale (ETS, 1976). The non-symmetric scale is designed to limit the effect of raters' tendencies to under-estimate item difficulty, but there is debate about whether this modification is appropriate. Other modifications include the use of multiple iterations (Melican & Mills, 1987; Cross, Impara, Frary, & Jaeger, 1984) and incorporation of ratings of item relevance.

The Ebel Method. The Ebel method requires an additional type of judgment about test questions. Items are rated on both their difficulty (easy, moderate, or hard) and relevance (essential, important, acceptable, or questionable). The ratings
are used to place items into a 3 X 4 matrix. Next, raters estimate the percentage of items in each cell that will be answered correctly by the minimally competent examinee. The standard is calculated by multiplying the number of items in each cell by the proportion of items the minimally competent examinee is expected to answer correctly and summing the values. Variations on the method involve modifying the values of the relevance scale (Garvue et al., 1983; Skakun & Kling, 1980) or using a different scale, for example, item importance (Cangelosi, 1984; Skakun & Kling, 1980).

An example of one rater’s application of the Ebel method is shown in the four panels in Figure 4. The top panel shows the rater’s placement of items into the cells in the matrix. Items 1, 8, and 15, for example, have been rated as easy and essential. The next panel contains the count of items in each cell. The rater’s predictions of the proportion of items in each cell that will be answered correctly by the minimally competent examinee are shown in the third panel. The values for each cell in the last panel are calculated by multiplying the number of items in each cell (the second panel) by the predicted performance for that cell (the third panel) The products are summed to produce a cut score.

An advantage of the Ebel method is that raters explicitly evaluate each item not only on its difficulty, but also on its relevance. Rating items on both dimensions allows hard, but essential, items to receive a higher rating than hard items of more questionable relevance. This provides raters the opportunity to adjust explicitly their expectations of performance based upon their evaluation of the appropriateness of the test content. (This practice could be viewed as inappropriate because, presumably, test content is based on a job analysis or similar procedure and all content is, therefore, presumed to be relevant.) Predictions of the expected performance of minimally competent examinees are based on groups of items, not
### Placement of 15 Items into Categories by One Judge

<table>
<thead>
<tr>
<th></th>
<th>Essential</th>
<th>Very Important</th>
<th>Important</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>1, 8, 13</td>
<td>2, 14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>9</td>
<td>3, 6, 12</td>
<td>7, 11</td>
<td></td>
</tr>
<tr>
<td>Hard</td>
<td></td>
<td>15</td>
<td>4, 5, 10,</td>
<td></td>
</tr>
</tbody>
</table>

### Number of Items Per Category

<table>
<thead>
<tr>
<th></th>
<th>Essential</th>
<th>Very Important</th>
<th>Important</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Hard</td>
<td></td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

### Predicted Proportion Correct by Category

<table>
<thead>
<tr>
<th></th>
<th>Essential</th>
<th>Very Important</th>
<th>Important</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>0.95</td>
<td>0.85</td>
<td>0.75</td>
<td>0.50</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.90</td>
<td>0.80</td>
<td>0.60</td>
<td>0.30</td>
</tr>
<tr>
<td>Hard</td>
<td>0.75</td>
<td>0.60</td>
<td>0.45</td>
<td>0.15</td>
</tr>
</tbody>
</table>

### Cutoff by Category and for Total Test

<table>
<thead>
<tr>
<th></th>
<th>Essential</th>
<th>Very Important</th>
<th>Important</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>2.85</td>
<td>1.70</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.90</td>
<td>2.40</td>
<td>0.00</td>
<td>0.60</td>
</tr>
<tr>
<td>Hard</td>
<td>0.00</td>
<td>0.60</td>
<td>1.35</td>
<td>0.00</td>
</tr>
<tr>
<td>Total Test Cut Score</td>
<td></td>
<td></td>
<td></td>
<td>10.4</td>
</tr>
</tbody>
</table>

Figure 4. An Example of the Ebel Method for 1 Rater and 15 Items

Individual items, which may be more accurate than predicting individual item performance. No research has been conducted, however, to verify this assumption. The requirement that judges perform multiple rating tasks makes training of judges, collection of data, and analysis of the data more complex than for other methods.

**The Jaeger Method.** The Jaeger method differs from other methods in the class in several ways. It incorporates ancillary information about the ratings of other experts and the impact of the ratings on passing rates in an iterative data collection design. The concept of the minimally competent examinee is not explicitly used. The item rating is based on a judgment about the importance of the item in relation to the decision to be made (e.g., “Should every beginning practitioner be able to

---

6Since the introduction of the Jaeger method, the provision of ancillary information (e.g., data on the ratings of other raters, impact of the ratings on passing rates) in iterative procedures with other standard setting methods has increased.
Should Every Beginning Practitioner Answer This Item Correctly?

<table>
<thead>
<tr>
<th>Item</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>8</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>10</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Cut Score</td>
<td>6.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. An Example of the Jaeger Method for 5 Judges and 10 Items

answer this item correctly?”). Selection of raters is not limited to individuals with subject matter expertise.

Initial standards are established by counting the number of items for which raters provide an affirmative response. Following the initial ratings, judges may revise their ratings after reviewing their cut scores, those of other judges, and the resulting passing rates.

Figure 5 contains an example of the initial ratings provided by five raters using the Jaeger method on a 10 item test. All raters agree that Item 1 should be answered correctly by beginning practitioners and all except Rater 5 agree that Item 10 need not be. Individual standards range from five to seven items correct with an average cut score of six items answered correctly.

Because the Jaeger method focuses more on an evaluation of test content than the interaction of the minimally competent examinee with test content, the standard setting process can include individuals who have an interest in the test results and content expertise, but who lack the familiarity with the examinee group necessary to focus on the minimally competent examinees only. Raters should, however, have sufficient experience with entry-level practitioners to be able to evaluate the test content relative to realistic expectations of the performance of those individuals. A potential drawback is that the rating task implies that passing status could be denied on the basis of an answer to a single item even though this is not how the method is implemented. Also, there is no clear rationale for how feedback about the expected pass rate or the test scores recommended by other raters should lead to revisions to individual item ratings.
10. ESTABLISHING PASSING STANDARDS

Summary. Methods based on the evaluation of test content are popular. Among the advantages of the methods are (a) cut scores can be estimated prior to the administration of tests, (b) familiarity with groups of examinees (not specific individuals) is the basis upon which judgments are made, and (c) the rating tasks tend to be straightforward. However, the methods also have drawbacks. Estimating performance on individual items is difficult. Most raters are not able to estimate item level performance with great accuracy (Lorge & Kruglov, 1953; Thorndike, 1982; and Bejar, 1983). Another drawback of the methods is that they do not provide data on expected pass rates or misclassification errors. There is no way to evaluate the results of the individual judgments to determine their “accuracy.”

Absolute Methods Based on the Evaluation of Individuals

Standard setting methods in this class rely on judgments of the expected passing status of individuals. Cut scores are established to maximize the agreement (typically) between the examinees’ expected passing status and the observed test scores. The best known methods in this category are the contrasting groups and borderline group methods (Livingston & Zieky, 1982).

Contrasting groups. The contrasting groups method requires score distributions for two groups of examinees: those expected to pass (competent) and those expected not to pass (not competent). Judgments of who is expected to pass and who is not expected to pass are typically made by the instructors who have trained the examinees. The method allows assessment of the number of classification errors (qualified individuals who fail and unqualified individuals who pass). Several assumptions are made about the method. First, the group of examinees at hand are representative of examinees who will be licensed using the same test. Second, the test will be used to make a decision about the group of examinees on hand and who have been classified as either competent or incompetent by their instructor and for future examinees who will not be classified by instructors or others independent from the test. Third, the more competent examinees will obtain higher scores on the test and the less competent examinees will obtain lower scores, but some examinees classified as competent will obtain low scores and some examinees classified as incompetent will obtain high scores.

To illustrate the contrasting groups method a data set was generated for a hypothetical sample of 342 examinees. Based on assumption one above, these examinees are assumed to be a representative sample of all examinees who will be licensed or not based on their score on the licensure examination. These data are shown in Table 2 (a graphical representation is shown in Figure 6). In this data set, 224 candidates were classified as competent (expected to pass) and 118 were classified as not competent (expected to pass).

In the contrasting group method the candidates are classified prior to testing (or, if after testing, without knowledge of the test score). After the test has been administered and scored, the distribution of examinee scores are partitioned at each score point into those examinees who were previously classified as competent and those who were classified as incompetent. The cut score is established by identifying the score that best represents the importance of the decision. That is,
if it is equally unacceptable to pass someone who should have failed as it is to fail someone who should have passed, the standard will be set at the score point where 50% of the examinees were classified as competent and 50% were classified as incompetent. In Table 2 this point corresponds to a score of 6. If passing an incompetent candidate was a more serious error (e.g., suppose it was considered twice as bad to license an unqualified candidate as to deny a license to a qualified candidate), then one might select the cut score such that the number of qualified who pass is twice that of the number of unqualified who pass. In Table 2 there is no passing score that corresponds exactly to that criterion, but the score of 9 comes closest (where 71% of those who scored a 9 were classified as competent, i.e., were expected to pass).

When using actual data, it may be the case that the distributions of scores for those expected to pass and those not expected to pass do not fit the assumptions above. Specifically, the scores of the examinees classified as competent do not
increase smoothly and the number of examinees classified as incompetent do not increase progressively at each lower score point. For this reason, Livingston and Zeiky (1982) have proposed techniques for smoothing the data (statistically adjusting the distributions) to accommodate the unevenness that might occur when dealing with real data, especially when the number of examinees is relatively small.

**Borderline Group.** The borderline group method basing the cut score on the test performance of individuals who have been independently designated as neither competent nor incompetent. The cut score is typically placed at the median of the scores of the borderline examinees. If, however, the consequences of the decision are such that the costs of passing individuals who are not qualified is unequal to the costs of failing those who are, a different placement of the cut-off score may be considered.

Figure 7 depicts the performance of 108 examinees classified as borderline on a 15-item test. The median of the group (i.e., the cut score) is at a score of 9.

A weakness of the method is that the number of examinees rated as borderline is often small. Thus, a cut score may be established using a small and possibly unstable distribution of scores. Furthermore, the distribution of scores for the borderline group overlaps with those of competent and not competent groups. As a result, a cut score that fails half the borderline group students is likely to be

---

Some experts object to the borderline group as being the wrong group upon which to base a cut score. Their argument is that the cut score should identify the minimally competent, not those who are neither competent nor incompetent. See Kane (1994) for a discussion of this issue.
different from one that best separates competent and not competent groups. In most situations, if it is possible to collect borderline group ratings, it will also be possible to collect data to implement the contrasting groups method. If so, contrasting groups is preferable because the data are directly related to the decision to be made (establishing a standard that separates competent from incompetent examinees).

**Setting Standards on Performance Assessments**

A recent trend in assessment is the inclusion of performance tasks in tests. Some of these tasks are relatively simple (e.g., writing an essay), but complex performance assessments are also gaining popularity. Complex performance assessments require examinees to perform tasks that have many components, each of which is important to job success. Such assessments are viewed as more relevant than the traditional multiple-choice tests that dominate most licensure tests.

Despite increasing use of performance assessments, there are many psychometric issues to be addressed. Issues such as topic selection, generalizability of the
results, and scoring methods are being actively researched. Similarly, little
guidance is available for establishing cut scores on complex performance assess­
ments although this area is also being actively researched. This section describes
some of the methods under investigation.

Simple Performance Assessments

In some cases a complex assessment may generate a simple result. For
example, a diagnostician might be given a series of laboratory results and be
required to write a report summarizing those results. A single score may be
generated to summarize the adequacy of the report. In such cases, an independent
group of raters (i.e., not the individuals who score the assessment) might read the
reports and classify them as acceptable, unacceptable, or borderline. The contrast­
ing groups or borderline group method can then be used to determine the cut score.
In these cases, many of the limitations of these methods are reduced because the
judgments are made on a work product, not on the individual. Thus, in the case of
a performance assessment that yields a single, summative score, the standard
setting task is relatively straightforward.

Complex Performance Assessments

In contrast to the simple example above, consider a laboratory assessment in
which the examinee is required to draw a sample, conduct tests using the sample,
and write a report. Several such tasks might be included in a single examination
so that different types of samples must be drawn using different equipment,
different analyses will be conducted, and several different types of reports may be
required (e.g., an internal report, a report for a third party, or a report to the patient).
As a result, there may be many tasks and each task may assess multiple (but not
necessarily all) dimensions of performance. Thus, there can be several types of
scores (in this example, scores within task, task scores, and a test score). Thus, the
assessment is multi dimensional and the standard setting process will need to take
this into account. An example of a complex performance assessment is shown in
Figure 8. The test consists of three tasks (A, B, and C). Five skills are assessed,
but not every skill is assessed for every task. Skills 1 through 4 are assessed on two
of the tasks, but Skill 5 is only assessed in Task B. Scores are generated on, for
example, a scale of 1 to 4 on each skill.

<table>
<thead>
<tr>
<th>Skill</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8. A Design for a Complex Performance Assessment
Three methods have been proposed for setting standards on assessments such as the one described above. These are two-stage judgmental policy capturing (Jaeger, 1994), extended Angoff (Hambleton & Plake, 1994), and multi-stage dominant profile analysis (Putnam, Pence, & Jaeger, 1994). The methods have not been used operationally and it is unclear how (or if) they will be implemented. That notwithstanding, they represent the current state-of-the-art and should be considered by test sponsors using complex assessments.

Two-stage Judgmental Policy Capturing. Judgmental policy capturing relies on regression analysis of raters’ judgments about profiles of scores to determine the standard. In the first stage, raters are shown profiles of scores on skills measured by each task. The raters judge the profile (e.g., Poor, Mediocre, Satisfactory, Noteworthy, Excellent). For the second stage, profiles are generated based on evaluations of the individual task ratings in the first stage. These profiles are then rated according to the decision to be made on the basis of the test results (e.g., Novice, Competent, Accomplished, and Highly Accomplished). Figure 9 contains

---

*These labels were used by Jaeger (1994) to collect judgments designed to identify superior performance. Different labels might be used in different settings.*
two examples of profiles. The first is a profile of skill scores on Task A and the second contains a profile for the three tasks.

**Extended Angoff.** Hambleton and Plake’s (1994) extended Angoff method is an extension of the Angoff method described earlier in this chapter. That is, raters provide their expectations of the score of a minimally competent examinee on each dimension for which scores are generated. The Angoff method is extended by allowing raters to weight the skills according to their perceptions of the relative importance of each skill. Cut scores are established by multiplying the ratings by the weights and summing the resultant values.

**Multi-Stage Dominant Profile.** The multi-stage dominant profile method (Putnam, Pence, & Jaeger, 1994) was implemented as part of the same study in which the two-stage policy capture analysis and the extended Angoff methods were introduced. It was developed in response to raters’ dissatisfaction with the other methods, especially the extended Angoff method. The method incorporates more direct data collection about raters’ policies regarding acceptable performance through a three-stage process: policy creation, feedback, and implicit policy generation.

**Policy Creation.** In this stage, raters generate profiles depicting their perceptions of acceptable performance. The profiles show scores on each skill within each task that, taken together, would be considered acceptable. Multiple profiles are generated to depict the variation in performance that can be considered acceptable. A written statement is generated summarizing the policies underlying the profiles.

**Feedback.** Raters review their profiles and the profiles of other experts. Additional profiles are generated and evaluated by the experts.

**Implicit Policy Generation.** A series of “challenge profiles” (profiles that reflect the policy statements in most, but not all ways) are generated and submitted to raters for a final evaluation. Raters judge these profiles with a simple Yes/No response to the question of whether or not the performance was acceptable. Final standards are generated through a logistic regression.

**Issues in Setting Standards on Complex Performance Assessments.** As noted above, the use of complex performance assessments is not yet widespread and there are many issues to be resolved before standards for professional practice emerge. However, as such assessments gain popularity, they will undoubtedly be used as part of the licensure process.

To date there are no established methods for setting standards on complex performance assessments. The methods that have been proposed are complex (conceptually, operationally, and analytically). Furthermore, the methods require raters to consider issues (such as the weighing of scores) that have not traditionally been part of the rating portion of a standard setting study. At this point, it is unclear whether the methods will be refined in ways that allow their routine use in licensure settings or whether other methods will have to be developed.

**WHICH METHOD IS BEST?**

None of the methods described above can be designated as the “best” because there is no way to verify their validity. However, Berk (1986) has listed criteria for
evaluating standard setting methods. Using prior research, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education Joint Committee, 1985) and court decisions, Berk developed technical and practicability criteria that can be applied to all methods. Drawing from Fitzpatrick (1984), he also listed additional criteria that apply specifically to methods based on evaluation of test content. Berk’s criteria are summarized below:

**Technical Criteria**

1. The method should classify test takers into mutually exclusive groups.
2. The method should be sensitive to the difficulty of the test.
3. The method should incorporate evaluation of the opportunities examinees have had to learn the material presented (unless that information is gathered elsewhere).  
4. The method should yield appropriate statistical information.
5. The method should take into account differences between the “true” standard (on the true-score scale) and the observed standard.
6. The method should allow for evaluation of classification errors.

**Practicability Criteria**

7. The method should be easy to implement.
8. The results should be easy to compute.
9. The explanation of the method should be understandable by people who are not experts in measurement.
10. The method should be credible.

**Additional Criteria**

11. The effect of “social comparisons” (raters comparing themselves to other, more influential raters) should be minimized and informational influences maximized.
12. Exposure to the opinions of others can result in raters changing their views to conform to the opinions of others and should be avoided.
13. Group discussion among the raters is desirable, but is likely to be biased in favor of the majority opinion of the group unless structured procedures are implemented to ensure that all positions are stated.
14. The effect of normative judgments about ratings can be limited by providing objective information about test performance.
15. If opportunities are provided for revision to judgments, public statements of initial positions should be avoided.

No method satisfies all of the criteria. Depending on the situation, however, any of the methods described in this chapter can yield an acceptable and defensible cut score. However, consideration of these criteria in conjunction with other information (e.g. the importance of the decision, political considerations in the process, etc.) can help guide the selection of the most appropriate method for a given situation.

---

9 An argument can be made that this criterion should not apply to licensure tests. If the content of a question covers a critical component of the professional that is required to protect the public, opportunity to learn may be relatively unimportant to the licensure decision.
Planning Study Procedures and Analyses

Data collection procedures in standard setting studies can be complex. Among the decisions that will affect procedures are whether (a) multiple iterations will occur; (b) feedback will be provided about the ratings of others; (c) raters will discuss their ratings and, if so, at what point; (d) limits will be placed on judgments of item difficulty (e.g., use of discrete categories with the Angoff method); (e) corrections for guessing will be applied to the ratings or the resulting standard; and (f) actual test or item performance information will be provided during the ratings and, if so, at what points. Another decision that will affect procedures is the timing of the study (before or after the test has been administered). Procedures should also address how item criticisms will be handled. Each issue is summarized briefly below.

MULTIPLE ITERATIONS

Some standard setting studies involve multiple iterations. Following an initial rating, additional information is provided. This information can consist of summaries of the ratings of individual judgments, data on item performance, information on the effect of the initial ratings on passing rates, and so on. The exact information provided depends on the design of the study and the data available. In some cases, there are two iterations and in others, three iterations occur.

The Jaeger method incorporates iterative judgments into the process. Following the introduction of the Jaeger method, iterative procedures became more popular with other methods as well (Mills & Melican, 1990; Melican & Mills, 1987; Cross, Impara, Frary, & Jaeger, 1984). Given the increased interest in providing feedback to raters, iterative procedures are gaining acceptance. However, use of an iterative procedure assumes the capability to summarize ratings in a standardized manner as the study progresses (i.e., on a “real-time” basis). If this cannot be done, the value of an iterative procedure is lessened although group discussion (based, for example, on a show of hands about item ratings) is a useful method for providing a basis for revising initial ratings.

PROVIDING FEEDBACK ON THE RATINGS

A common feature in iterative procedures is the provision of information to the raters on the ratings provided by others. Research indicates that providing information on the ratings of other experts often results in revisions to initial ratings (Friedman & Ho, 1990; Busch & Jaeger, 1990; Melican & Mills, 1987). Typically the number and magnitude of revisions is small. However, the studies suggest that the revisions usually result in reduced variation across judges and increased accuracy with regard to actual item difficulty.

DISCUSSION OF RATINGS

Allowing judges to discuss their ratings, identify items for which there is significant variation among the ratings, and determine items that one or more raters may have misinterpreted is common. Most iterative procedures provide for group discussion of ratings. The timing and extent of the discussion vary. Some investigators allowing discussion during the initial rating, some during the second
iteration (e.g., Busch & Jaeger, 1990), and some following the second iteration (Melican & Mills, 1990). Given the advice of Fitzpatrick (1984) that group discussion can have negative influences, discussion during the first rating is probably undesirable. Even when discussion occurs following the initial rating, the investigator should ensure that the discussion is structured in such a way that a single individual cannot dominate and that all raters have opportunities to provide input to the discussion.

PLACING LIMITS ON THE JUDGMENTS

Some investigators place lower limits (i.e., chance) on the ratings provided. However, some items are quite difficult and particular distractors may be appealing to individuals with partial knowledge, so it is not uncommon for examinees to score below chance on those items. Therefore, this practice is not recommended. Reid (1985) investigated the effect of placing upper limits on ratings. Raters first estimated item difficulty for the total group and then estimated difficulty in the minimally competent group. This procedure resulted in lower ratings than a control group, which rated item difficulty for the minimally competent group twice. Reid concluded, however, that the results were inconclusive as to whether the procedure resulted in more “realistic” estimates. Neither placement of lower or upper bounds on ratings has been widely used.

ADJUSTING RATINGS FOR GUESSING

Some investigators apply corrections for guessing to estimates of item difficulty. If a test is scored using a penalty for incorrect answers, each rater’s cut score may be adjusted downward to correct for this penalty (Livingston & Zieky, 1982). Cross et al. (1984) point out that the wording of the task assigned to the raters can alleviate the need for corrections for guessing. Asking the judges to estimate what examinees would do incorporates guessing behavior in the estimates. Asking what the minimally competent examinee would know does not incorporate guessing and provides a statistically appropriate basis for making a correction for guessing. Melican and Plake (1984) point out, however, that this adjustment, which raises the standard, may be overly harsh if examinees omit questions.

PROVIDING FEEDBACK ON EXAMINEE PERFORMANCE

When standards are to be set on existing tests for which performance data (item difficulty and score distributions) are available, the data can be provided during the standard setting study. This can serve to set an upper limit on ratings, but unlike the Reid (1985) procedure, the data are from examinees, not from raters’ previous estimates of performance. Norcini, Shea, and Kanya (1988) and Melican and Mills (1986) recommend this procedure as one that can improve the accuracy and consistency of ratings. Some investigators have also attempted to use performance data to calibrate ratings as a means of equating (Rogosa, 1982; Thorndike, 1982).

---

10 If the test is scored on the basis of the number of questions answered correctly, corrections for guessing may still apply. If raters estimate what an examinee would know, the estimates do not include the number of questions that would be answered correctly due to guessing.
TIMING OF THE RATINGS

Typically, item ratings are collected in a special study using an intact test form. An expert group is convened and trained. The experts then provide estimates in one meeting. One study (Norcini, Lipner, Langdon, & Strecker, 1987) suggests that it may be possible to conduct the ratings by mail. This study is limited, however, in that the same raters provided three sets of ratings (before, during, and after the meeting) and the raters were the same individuals who wrote the test questions.

Ratings can be provided when the questions are written, when they are reviewed, when pretest data have been collected, or immediately following the first administration of the test. If a study incorporates feedback to raters on examinee performance, ratings need to occur following either pretesting or the first administration. Using pretest data is appealing because the data can be used to establish a cut score before the test is administered (examinees then know the “rules of the game before they play”). Care should be taken, however, to ensure that the pretest data are reliable. If examinees are aware that the pretest does not count, there is a risk that they will not take the test seriously and pretest statistics will indicate that the questions are more difficult than they really are.

ITEM CRITICISMS

It is not unusual for experts to object to the wording or key of a question during the rating session, especially if test development committees (who typically provide item reviews as part of their work) provide the judgments. It is important to recognize that, although items may need additional reviews and revisions, they are presented to judges under the assumption that they are of sufficient quality to be administered to examinees in their current form. Thus, ratings should be provided on the items as presented. A mechanism should be available, however, to allow experts to register their concerns and suggest item revisions. This will allow raters to identify items for further review (or discussion following the rating) without distracting them from the task at hand. Ratings can be gathered on the original and revised version of the item and, following a decision about which version will appear in the test, the appropriate ratings can be used to derive the standard.

Initial study results are often modified (see the section “Adjusting the Standard” below), so decisions are required at this stage concerning which method will be used to modify the study results and, if the method relies on expert ratings, forms will be required to collect those data.

Select Expert Reviewers

Virtually all standard setting methods require input from experts. Not all members of a profession will be qualified to be raters and different methods may require experts with different experience. Experts will need specific knowledge, skills, and experiences for the tasks they are to perform. The selection process should ensure, to the extent possible, that experts represent the full diversity of the profession and the various constituencies affected by the test. A more complete discussion of the qualifications of expert raters can be found in Jaeger (1991).

A typical question that arises is how many judges are required. The usual answer is “as many as can be obtained,” but this provides little practical guidance.
Norcini, Shea, and Grosso (1991) argue that acceptable results can be achieved with as few as five raters. Jaeger (1991), however, recommends calculating the number of raters based on the standard error of the mean of the ratings and the standard error of measurement of the test. Jaeger’s work suggests that the number of raters should be substantially greater than Norcini et al. recommend. The exact number depends on the precision desired. In one example, Jaeger’s procedures would require 13 raters to obtain a standard error in the ratings that is one quarter the standard error of measurement of the test.

**CONDUCT THE STUDY**

If the study design and planning have been comprehensive, there will be sufficient staff, materials, and equipment on-site for the study. Thus, the mechanics of data collection, form design, etc. are not discussed here. However, an important component of the study is the initial training of the raters. The training session, held prior to the actual rating of test items, frequently consists of four components: explaining the process, setting the context of the task, developing a common definition of the minimally competent examinee, and training judges to rate items.

Although the standard setting literature indicates that training is important (Mills & Melican, 1988; Fitzpatrick, 1984; Livingston & Zieky, 1982) little documentation is available regarding specific approaches to training. Much of the available literature addresses training in the context of applying the relevant procedure, not training related to defining minimal competence (Mills & Melican, 1986; Francis & Holmes, 1983).

An approach to developing a definition of minimal competence was proposed by Mills, Melican, and Ahluwalia (1991). The approach relies on group discussion to establish the definition of minimal competency and requires a substantial time commitment.

**Introductory Session**

Most raters will not have previously participated in a standard setting study and are unlikely to be familiar with standard setting techniques. An introductory session that provides an overview of the process, their roles, the data collection forms and use of the data can minimize confusion later. Raters will vary in their knowledge of test content, the purpose of the test, and the overall licensure process. They are also likely to vary in their support for the use of the test in the licensure process. The initial session should address these issues to reduce the probability that the ratings will be affected.

An understanding of the decision to be made on the basis of the test results is important. If, for example, the test is an assessment of academic knowledge, raters need to understand that predicting on-the-job performance is not of concern. Raters should understand that the test will not assess every aspect of the job and that their task is not to critique the test or its content, but rather it is to estimate performance on the instrument as it exists. Knowledge necessary to protect the public is an appropriate focus. Raters frequently have concerns about test content, the adequacy of content coverage, and test format. These are important concerns, however, they have usually been addressed separately as part of the test development process.
Typically, for licensure examinations, a job analysis will have been conducted and the test content specifications will be the basis for the content specifications.

A brief discussion of the development of the content specifications and test items can address these concerns and reduce their effect on the ratings. As a result, raters should understand what work has occurred prior to the study and their role in the overall process.

Defining Minimal Competence

An explicit definition of minimal competence is required for most standard setting procedures. Simply, minimal competence is the “minimal level of knowledge and skills required for licensure.” Unfortunately, this simple definition is not an operational definition of minimal competence and, therefore, is inadequate given the variety of skills being tested, the different ways they can be acquired, and the possible compensatory effects that strengths in one area might have for weaknesses in another area.

A discussion of minimal competence may begin by delineating the skills routinely required in practice. Refinements can then address typical and minimally acceptable proficiency (such as common, but acceptable errors). Using the test specifications can limit the discussion to those skills assessed by the test. Each major area of the specifications should be discussed.

The initial discussion about the range of skills in the general population of practitioners can be refined to focus on the level of those skills required for licensure. For example, inefficient procedures may not represent good practice, but they may be acceptable when the focus is the granting of a license. Statements of typical proficiency should be refined further to apply directly to the granting or renewal of a license.

The purpose of the discussion is to develop a concise definition of minimal competence. When completed, it may address the following statements:

A minimally competent examinee must know AT LEAST ...

A minimally competent examinee would not be expected to ...

The purposes of the training are to (a) set the context within which the ratings can occur; (b) define the tasks to be performed (and those not to be performed) by the raters; (c) eliminate, to the extent possible, the effect of irrelevant variables from the rating session; and (d) develop a common definition of minimal competence. The goal is not to have agreement on all ratings, but to ensure that differences are not due to irrelevant factors.

Training the Raters

Following the establishment of a definition of minimal competence, but prior to the actual ratings, a training session should be held to ensure that raters understand the rating task and have some understanding of the difficulty of the questions to be rated. The need for training is evident when the literature on accuracy of item ratings is reviewed. Numerous studies have documented the tendency of judges to under-estimate item difficulty and to achieve only modest correlations between actual and estimated difficulties (Lorge & Kruglov, 1953; Halpin & Halpin, 1983; Bejar, 1983; Thorndike, 1982; Schaeffer & Collins, 1984).
However, as noted previously, provision of information on item difficulty can improve ratings. Therefore, training which allows raters to compare their estimates with actual data is appropriate (Mills & Melican, 1986). The training should also include practice on all types of items included in the test because research has also shown that certain characteristics of questions can make them more difficult to rate accurately (Melican & Thomas, 1984; Smith & Smith, 1988).

There is no generally established guideline for how extensive the training should be. However, Reid (1991) has proposed three criteria for determining whether raters are well trained. According to Reid, ratings should be stable over time, consistent with the relative difficulties of items, and realistic relative to actual performance. Saunders and Mappus (1984) suggest that final results may be more consistent, accurate, and homogeneous if the results from raters who do not meet training criteria are eliminated from the analysis. Care should be taken in doing so, however, because the representativeness of the group may be threatened (Reid, 1991) and it is not necessarily the goal of a standard setting study to reduce variations in the ratings. Furthermore, unless the criteria for exclusion are established prior to the study, criticisms may be raised about the appropriateness of the procedure.

EVALUATE RESULTS AND ESTABLISH STANDARDS

The results of the study should be carefully reviewed to ensure that the experts understood the task, were diligent in their application of study procedures, and that the procedures established were adequate. A careful review of the results can identify flaws in the study that may possibly be corrected. In some cases, this pre-analysis will lead to the conclusion that the study must be repeated. Although it is unpalatable to repeat a study, there are occasions when this is the only feasible solution. For example, in some cases, it will become clear that most raters did not understand their assigned tasks. In this case, there is no way to use the study results appropriately and new panels must be convened.

Many factors can (and should) be considered in the establishment of the final standard. The standard setting data are of great importance and value; however, it should be remembered that this information was provided in a very specific setting, focusing (usually) on only the content of the test or the test and the examinees taking it. It is not unusual for test sponsors to carefully plan a standard setting study, but to ignore the need to consider the results of that study in the context in which it will be applied. For example, a standard that is too stringent could result in serious shortages of licensed professionals, whereas one that is too lenient could put the public at risk. In either case, the entire testing program could be called into question. Therefore, planning should include consideration of how the cut score derived from the study will be evaluated and, if necessary, adjusted, and by whom.

Adjusting the Standard

The test sponsor’s governing board or council typically has the ultimate responsibility for establishing the standard. If the board adequately represents all interested constituents, it may receive the study results directly and establish the standard. However, sponsors often wish to include others in the evaluation of the
study results before establishing the standard. For example, a review group composed of representatives from employer organizations may be convened to review the study results and recommend the final standard. Ultimately, however, the final decision rests with the test sponsor or legal authority charged with establishing the standard. Both the advisory panel and the decision makers will need to consider whether or not the study results require adjustment. Several methods of adjusting the standard are available.

**Standard Error of Measurement Adjustment**

The standard error of measurement is an estimate of the dispersion of individuals’ observed scores around their true scores (see Chapter 7, Impara & Stoker for a more extensive discussion). Errors of measurement can result in two types of classification errors. Individuals whose true score is just above the cut score may fail because errors of measurement result in an observed score that is lower than both the true score and the cut score. Lowering the cut score by a multiple of the standard error of measurement decreases the likelihood of this type of error. However, it increases the likelihood that individuals whose true score is slightly below the cut score will pass. The method is implemented by considering (a) the relative seriousness of each type of classification error and (b) the effect of measurement error on scores near the cut score. For example, if it is worse to fail a qualified individual than to pass an unqualified one, one standard error of measurement might be subtracted from the study value. Adjusting for errors of measurement is a common and defensible method for establishing cut scores.\(^{11}\)

There are, however, disadvantages to the standard error of measurement adjustment. It assumes that the cut score derived from the study is “correct” and that the only adjustments required are those necessary to account for measurement errors. Furthermore, discussions about the relative costs of the two types of error are sometimes in contradiction to the test results. It is not unusual for test sponsors to state initially that passing someone who should fail is worse than failing someone who should pass. Using the standard error of measurement adjustment, this would lead to a decision to raise the cut score. However, in practice, raters’ expectations of performance often exceed actual performance and result in a need to lower the cut score (not due to errors of measurement, but due to overly optimistic ratings). Although the standard error of measurement adjustment should be a philosophical one that does not rely on test data, decision makers are often reluctant to make the decision without information about the impact of the adjustment. Although pass rates are useful in assessing the reasonableness of a cut score, the standard error of measurement adjustment can be criticized if it appears to have been used solely to adjust the pass rate without regard for the philosophical basis for the adjustment. However, methods for directly incorporating ratings of expected passing rates have been proposed and are described in the next section.

\(^{11}\) It has also been suggested that cut scores might be adjusted using the standard error of the judges. This treats the raters as a random sample of potential panelists and the cut score is adjusted to compensate for possible sampling error. The method suffers from many of the same drawbacks as the standard error of measurement adjustment.
Observed Score Distribution Adjustments

The knowledge that experts have about the examinee population and their expectations about the percentage of the population that will pass the test can be used in conjunction with other information to establish final standards. The methods proposed by Beuk (1984), De Gruijter (1985), and Hofstee (1983) incorporate judgments about the expected performance of examinees. These methods assume that experts can provide estimates of the passing rate. All of the methods require observed score distributions, therefore, they cannot be implemented prior to the administration of the test (although the data can be collected prior to the test administration and applied before scores are reported). One of the strengths of these methods is that because the data used to calculate the adjustment are collected without knowledge of score distributions, they are less susceptible to criticisms that the standard was arbitrarily adjusted to yield an acceptable pass rate.

The Beuk Method

The Beuk (1984) method requires an estimated cut score and passing rate from each rater. The adjustment is a function of the variability of the experts’ estimates of the cut score and passing rate. To implement the method, a line with slope equal to the ratio of the standard deviations of the experts’ estimates of the cut score and passing rate is drawn through a point defined by the average absolute cut score and average passing rate. The intersection of this line and the cumulative frequency distribution becomes the recommended cut score. An example of the method is shown in Figure 10.

![Figure 10. An Example of the Beuk Method](image-url)
The Beuk method is straightforward. The only data required are a frequency distribution of test scores, the estimated cut scores and the expected passing rates. The computations are simple and the adjustment is logical; the more the judges agree on their estimates on one dimension (i.e., cut score or passing rate) the smaller the adjustment on that dimension. In practice, however, some experts have difficulty specifying expected passing rates if they have not had experience with large numbers of newly licensed practitioners.

The Hofstee Method

The Hofstee method requires estimates of the highest and lowest acceptable cut scores and passing rates. Two points are plotted using these four numbers. One point is defined by the minimum acceptable cut score and the maximum acceptable fail rate. The maximum acceptable cutoff score and minimum acceptable fail rate define the second point. Any point falling on the line segment defined by these two points is considered an acceptable combination of cut score and failing rate. The intersection of the line segment with the cumulative frequency distribution of scores defines the cutoff score. An example of the Hofstee method is shown in Figure 11.

The method is not complex. However, in practice the method is not always effective. The line segment depicting acceptable cut scores for any judge may not intersect the cumulative frequency distribution (Mills & Melican, 1987). In this case, the method cannot be used to adjust the standard because there is no acceptable combination of cut score and pass rate. (See the line for Judge 1 in

Figure 11. An Example of the Hofstee Method
Figure 11.) There are also questions about the global judgments of cut scores. If, for example, an Angoff method has been used, and separate estimates of minimum and maximum acceptable cut scores are collected, the calculated Angoff cut will not necessarily lie within the specified range of acceptable cuts.

The De Grujiter Method

The De Grujiter (1985) method is similar to the Beuk method. However, the De Grujiter method bases the adjustment on individuals’ uncertainty about the accuracy of their own ratings. After the raters have provided estimates of the cut score and expected passing rate, they also provide estimates of their uncertainty of the accuracy of their estimates. The adjustment is a function of the ratio of these uncertainty estimates.

The uncertainty ratings are the strength of the method. It is the only method that incorporates raters’ confidence in their ratings. The method is, however, computationally complex and difficult to explain. Further, experts frequently have difficulty specifying their uncertainty (Mills & Melican, 1987). Figure 12 shows an example of the DeGrujiter method.

OTHER FACTORS THAT MAY BE CONSIDERED

In addition to consideration of test reliability, estimates of test difficulty, and expected passing rates, there are other factors that may result in adjustments to the

![Figure 12. An Example of the DeGrujiter Method](image-url)
standard. Geisinger (1991) has listed several types of supplemental information that may be considered. The supplemental information that may be considered includes:

*Organizational or Societal Needs*

If the number of individuals needed can be predicted accurately, the cut score can reflect this. It may be unreasonable to designate individuals as passing a test if they have little opportunity to be hired (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978).

*Adverse or Disparate Impact Data*

Consideration of passing rates for gender, race, and ethnic subgroups should be considered. This topic is covered in depth in Chapter 2.

*Anomalies in the Rating Process*

In the course of the study (or evaluation of the results) it may become apparent that there were problems with the evaluations provided by the judges. These problems could result in elimination of one or more rater’s judgments or a new study. In other cases, the problems may require less severe remedies. Possible problems include (a) individuals who are designated as experts may prove not to have sufficient expertise for the task, (b) one or more raters may have misunderstood the task, (c) personal stakes in the outcome of the test may affect a rater’s estimates, (d) some raters may be unduly influenced by others, (e) the group of raters may be insufficiently representative of the field, and (f) a rater has provided clearly inappropriate ratings (e.g., all items receiving the same rating). As noted previously, decisions to eliminate ratings should be based on previously enunciated criteria to avoid the appearance of arbitrary manipulation of the results.

*Opportunities to Retest*

If tests are not offered frequently, failing the test may result in significant delays in the opportunity for entry to practice (upon taking and passing a subsequent test). Thus, it is especially important that individuals who fail the test are truly below the cut score. Within the bounds of other constraints (protection of the public, for example), a more lenient standard might be established if the opportunities for retesting are limited.

**MULTIPLE STANDARD SETTING TECHNIQUES YIELD DIFFERENT RESULTS**

On some occasions, a test sponsor may implement multiple standard setting methods or conduct multiple studies using a single method. Norcini and Shea (1992) and Mills and Melican (1990) have shown that consistent standards can be obtained across groups and occasions using the same method. However, it is equally clear that different methods yield different results (see Jaeger, 1989). If multiple methods are used, a rationale will be required for choosing the method implemented or for the manner in which the results from the different methods are combined to arrive at a final standard.

---

12This issue is not relevant to licensure tests, but is included for completeness.
Establishing a cut score defines an arbitrary division of a continuous variable into a dichotomy. It does not represent “truth,” but rather it is a representation of the collected wisdom (values) of professionals concerning the minimum skills necessary to enter the profession. Furthermore, because tests measure only a portion of what is important for success and because strengths in one area can often compensate for weaknesses in another, there will always be people who are qualified, but are denied licensure and some who are not qualified, but receive a license. This is not to say, however, that standards are indefensible. If a standard is developed based on the reasoned judgment of experts using a professionally accepted methodology, it can be defended. Comprehensive documentation of the study planning, procedures, and outcomes will play an important role in the event of a legal challenge.

All aspects of the process should be documented. Memos covering the planning process, the selection of experts, the actual study and the deliberations leading to the final standard should be included in the documentation. Samples of data collection forms should be retained as should the results of the analyses.

CONCLUSION

Establishing a passing standard is an integral part of most licensure programs. Despite years of research, there is still no one best method of setting a cut score that can be applied in all circumstances. However, there is a substantial body of research and practice that can guide the design and conduct of a standard setting study and the final establishment of the standard. This chapter has explained the steps in establishing a standard, reviewed methods available, and identified issues to be addressed during the process.

REFERENCES


