An NCME Instructional Module on

Setting Passing Scores

Gregory J. Cizek
University of Toledo

This module describes standard setting for achievement measures used in education, licensure, and certification. On completing the module, readers will be able to: describe what standard setting is, why it is necessary, what some of the purposes of standard setting are, and what professional guidelines apply to the design and conduct of a standard-setting procedure; differentiate among different models of standard setting; calculate a cutting score using various methods; identify appropriate sources of validity evidence and threats to the validity of a standard-setting procedure; and list some elements to be considered when evaluating the success of a standard-setting procedure. A self-test and annotated bibliography are provided at the end of the module. Teaching aids to accompany the module are available through NCME.

The term *standard* has a variety of related meanings: "standard equipment" on an automobile can actually mean little or no equipment; "standard accommodations" at a hotel implies a moderate level of luxury; a company's "standard reply" to a complaint connotes an apathetic response. At the extremes, a "substandard" manufactured part is defective; "holding up a standard" is associated with excellence.

However, limiting the use of the term to how it is most often used in educational measurement, a fairly consistent usage emerges. In this context, the term standard is usually shorthand for "standard of performance." Most often, to set a standard of performance means to implement a process that identifies a point on a score scale that divides the observed test score distribution, resulting in classifications such as master/nonmaster, pass/fail, or certify/deny certification. At other times, standard setting defines boundaries which define more than two states or degrees of performance, such as in the assignment of grades (e.g., A, B, C, D, F) or to differentiate between adjacent performance levels, such as in the achievement levels of basic, proficient, and advanced used on the National Assessment of Educational Progress (NAEP).

Gregory J. Cizek is an Associate Professor of Educational Research and Measurement at the University of Toledo, 350 Snyder Hall, Toledo, OH 43606-3390. His specializations are standard setting, applied measurement, and testing policy.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes. Professor George Engelhard, Jr., Emory University, has served as editor for this module.

A measurement specialization—standard setting—has developed to assist in deriving the level or levels of performance required. And, there are a wide variety of applications for standard-setting methods. Standards are established for kindergarten readiness; for student achievement in school subjects; for teacher and administrator proficiency in professional knowledge; for admission to institutions, programs, and services; for student placement and selection; and for candidates seeking certification and licensure.

Standard Setting: Perspectives and Definitions

Practically, *standard setting* is the process used to arrive at a passing score. The passing score is the lowest score that permits the examinee to be deemed competent, to receive a license or credential, to gain admission, and so on. Both Madaus (1992) and Zieky (1994) provide interesting histories of standard setting, tracing the policy uses and consequences of standards back over 2000 years. More recently, Linn (1994) has suggested that standard setting addresses four concerns: (a) exhortation, (b) exemplification, (c) accountability for educators, and (d) certification of student achievement.

Early standard setters frequently conceptualized the process of standard setting within the dominant paradigm of quantitative social science—that is, in the language and methods of estimating population parameters from sample observations. Jaeger (1989) reports that:

> Much of the early work on standard setting was based on the often unstated assumption that determination of a test standard parallels estimation of a population parameter; there is a "right answer," and it is the task of standard setting to find it. (p. 492)

However, by the late 1970s, measurement specialists had begun to debate whether setting standards could even legitimately be called a scientific enterprise. One frequently cited position in the debate was argued by Glass, who held that attempts to set standards were "either blatantly arbitrary or derive[d] from a set of arbitrary premises" (p. 258). He called the decision making process of standard setting "judgmental, capricious, and essentially unexamined" (1978, p. 253). Mostly, those who favored a position different than Glass's argued variously that standard setting was not an arbitrary process or, at least, that it was not arbitrary in the sense of being capricious (see Block, 1978; Popham, 1978).

The debate was as short as it was intense; standard-setting practice continued lacking a perceptible consensus on a theoretical foundation. However, an increasing number of measurement specialists began to reject the parameter estimation perspective as a framework for setting a standard. As Shepard has observed, "The standard we are groping to express is a psychological construct in the judges' minds" (1984, p. 188). Jaeger has also expressed the opinion that, "a right answer does not exist, except, perhaps, in the minds of those providing judgment" (1989, p. 492).

Two definitions of standard setting have begun to displace the parameter estimation perspective. Cizek (1993) has provided a *procedural* definition of standard setting that focuses on the process itself, employing an analogy to the legal concept of due process. He has defined standard setting as "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (p. 100). Cizek's definition eschews reference to a "true" cutting score that separates real, unique states on a continuous underlying trait (such as "minimal competence") and focuses instead on a process that can be used to rationally derive, consistently apply, and explicitly describe procedures by which inherently judgmental decisions must be made.

A second, *conceptual* definition is provided by Kane (1994a) who further refined the notion of standard setting and framed the process as a matter of score interpretation:

> It is useful to draw a distinction between the *passing score,* defined as a point on the score scale, and the *performance standard,* defined as the minimally adequate level of performance for some purpose. . . . The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version. (p. 426)

Guidelines for Standard Setting

Guidance for designing, conducting, and evaluating standard-setting procedures is available in the professional literature on the topic. Specific guidelines for standard setting are also provided in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1985). The *Standards* notes that: "defining the level of competence required for licensing or certification is one of the most important and difficult tasks facing those responsible for such programs" (p. 63).

The *Standards* contains several mentions of relevant standard-setting principles. Six specific references to standard setting are listed, with five of the six designated as *primary* and one guideline designated as *secondary.* The primary standards require standard setters to: describe how rates of misclassification will vary depending on the percentage of individuals tested who actually belong in each category (Standard 1.24); make information available regarding the rationale of the test and a summary of the evidence supporting intended interpretations, including information about the validity of the cut score (Standards 5.11, 8.6, 10.9); provide details on the standard-setting method used and the rationale for setting a cut score, including information about the qualifications of the participants in the process (Standard 6.9). The single secondary standard requires reporting of standard errors of measurement to be reported at critical score levels, especially at or near the cut score (Standard 2.10).

However, the 1985 version of the *Standards* also acknowledges that many developing areas in testing were not satisfactorily addressed, noting that: "These standards are concerned with a field that is evolving. Therefore, there is a continuing need for monitoring and revising this document as knowledge develops" (AERA/APA/NCME, 1985, p. 2). Currently, the *Standards* is undergoing its first revision in over a decade; perhaps additional professional guidelines on standard setting will be developed and incorporated.

Standard Setting: Preliminary Considerations

One bit of guidance that is not often considered is whether to have a test and/or passing score in the first place. Testing should not occur without a clear purpose or compelling reason. By extension, the implementation of a standard-setting method should be accompanied by sufficient justification. Kane (1994a, p. 427) has summarized the issue of purpose:

> Before embarking on any standard setting method, however, it is important to consider the fundamental issue of whether it is necessary or useful to employ a passing score. . . . Assuming that it is necessary or useful to employ a passing score, it is important to be clear about what we want to achieve in making pass/fail decisions, so that our goals can guide our choices at various stages in the standards-setting process.

One piece of guidance that cannot be stressed enough is the requirement for any standard-setting process to be carefully planned prior to implementation. The final desired product of a standard-setting procedure is, of course, a recommended passing score. The recommended score and details about how it was derived are usually presented to some entity which is actually responsible for the credentialing or licensure decision. However, the intermediate product of a standard-setting procedure is the standard-setting data. The quality of these data is probably the most important *measurement* consideration. As Cone and Foster observe, we can often fail to "evaluate whether the data [we] obtain so cleverly and analyze so complexly are any good in the first place" (1991, p. 653). Or, as others have noted: "You can't fix by analysis what you bungled by design" (Light, Singer, & Willett, 1990, p. viii).

Finally, it should be noted that standard-setting procedures are rarely an end in themselves. The cutting score that results from implementation of a standard-setting procedure is most accurately referred to as a *recommended* standard. Ultimately, the entity with the authority to establish a standard of performance must review, reject, adjust, or approve the results of the standard-setting process.

Standard-Setting Models and Methods

Early standard setting often utilized norm-referenced methods; occasionally these methods are used today. Norm-referenced or *relative* methods were described by Nedelsky as defining adequate achievement by a student "relative to his class or to any other particular group of students" (1954, p. 3). For example, a norm-referenced procedure for a credentialing examination might establish a passing mark at one standard deviation below the mean score for the group.

By the 1970s, with the proliferation of criterion-referenced testing, relative methods for setting standards were displaced by so-called *absolute* methods. Actually, the term *absolute* method had been suggested decades earlier by Nedelsky (1954) who sought a method of standard setting that would not be referenced to the performance of other students' performance on an examination but solely "on the instructor's judgment of what constitutes an adequate achievement on the part of a student. . . . In that sense the standard to be used for determining the passing score is absolute" (1954, p. 3).

Many of the absolute methods in use today were developed during a time Zieky termed an "Age of Awakening" in his history of standard-setting practice (1994, p. 10). These methods have subsequently been classified as either *state* or *continuum* models by Meskauskas (1976). State models assume that student competency is a truly dichotomous variable; continuum models view competence as a continuous variable, the distribution of which is artificially dichotomized by application of the cutting score. Because they "have not found wide applicability in competency testing programs" (Jaeger, 1989, p. 493), these models will not be explicated here.

On the other hand, continuum models have seen comparatively broader use in educational contexts, and Jaeger recommends that "in most practical applications, then, the choice of a standard-setting method should be restricted to . . . continuum models (1989, p. 493). Jaeger has suggested a further subdivision of continuum models into "test-centered" and "examinee-centered" (p. 493) to reflect whether judgments about competence are based primarily on inspection of test items (the test-centered models) or on judgments about examinees (examinee-centered models). However, as will be seen, both methods rely extensively on both kinds of judgments.

It should be noted that the test-centered and examinee-centered methods have additional characteristics in common. For example, both methods require judgments to be made by a group of persons qualified to make the judgments. (These persons are often referred to as *judges* in the literature, though *standard-setting participants* will be used here.)

A second commonality is that all standard setting requires some reference to a hypothetical person, such as the "minimally competent student," or to an abstraction in terms of performance, such as "borderline proficiency." Accordingly, a second key beginning question in all standard setting is how to train participants so that they acquire common conceptualizations of these critical reference points. Some suggestions for this training include selecting participants who are expert in the content area assessed and familiar with the ability of examinees in the range tested. Frequently, standard-setting participants are administered (and receive scores on) a form of the examination that will be used to make the certification, licensure, or mastery decisions. Other suggestions for helping participants to develop common conceptualizations and for training them to generate reasonable judgments are provided in Mills, Melican, and Ahluwalia (1991) and Reid (1991).

The fact that all standard setting requires human judgment leads to three key beginning questions in all standard setting: (a) Who should participate in the standard-setting process? (b) How many should participate in generating the judgments? (c) How should the judgments obtained from participants be evaluated?

The first question could be answered many ways depending on the kinds of judgments to be made, political considerations, and the purpose of setting a standard in the first place. In the context of student competency testing, Jaeger (1991) has provided guidelines for the selection of participants in the standard-setting process. He suggests "the ideal situation as sampling all populations that have a legitimate interest in the outcomes" (1989, p. 494), while others have recommended that participation be limited to those with expertise in the area tested. A reasonable synthesis of these bits of advice is to empanel standard-setting groups consisting of stakeholders with relevant knowledge of the area and population assessed.

The second question is, perhaps, easier to answer: The larger the panel, (usually) the smaller the standard error of the mean recommended standard. This principle is illustrated in Jaeger (1991) who provides the familiar formula for the standard error of the mean; however, Jaeger also rightly notes that the formula applies when "[a population of] judges are sampled randomly, independently, and with equal probability" (p. 5), conditions which are frequently violated in assembling standard-setting panels. In most cases, the most sensible guiding principle is simply this: Utilize as many participants as practical, given available resources.

The third questions is, perhaps, the most difficult to answer. As Engelhard and Cramer (in press) note, "relatively little research has focused on the general problem of how to evaluate the judgments obtained from standard-setting judges." The question is commonly addressed by examining the correlation matrix of participants' ratings, by examining correspondence with empirical *p* values, or by looking for round-to-round reduction in variability of participants' ratings. Engelhard and Cramer illustrate one method of evaluating the "goodness" of the data generated by the participants within an item response theory framework. Harnish and Linn (1981) have examined other *caution indices.*

Test-Centered Models

Just as criterion-referenced testing has proliferated, so too have the number of standard-setting methods available for these instruments. By the latter part of the 1980s, Berk had catalogued 38 alternatives in a "consumer's guide to setting performance standards" (1986, p. 137). Of these, however, perhaps three or four account for nearly all of current applications. The following sections provide a brief overview of the Nedelsky (1954), Ebel (1972), Angoff (1971), and Jaeger (1982) methods. Additional practical information on how to implement many of the procedures discussed in this section and the following section on examinee-based methods can be found in Livingston and Zieky's handbook called *Passing Scores* (1982) which continues to be the primary practical reference work on standard-setting procedures.

Nedelsky's method. Nedelsky's (1954) method involves assigning values to multiple-choice test items based on the likelihood of examinees' being able to rule out incorrect options. Nedelsky suggested the conceptualization of the hypothetical, minimally competent "F–D student" to assist in deriving a passing score. According to Nedelsky, on an individual item,

> Responses which the lowest D student should be able to reject as incorrect, and which therefore should be attractive to [failing students] are called F-responses.... Students who possess just enough knowledge to reject F-responses and must choose among the remaining responses at random are called F–D students. (1954, p. 5)

To use the Nedelsky method, standard-setting participants carefully inspect test items and identify, for each item in the test, any options that a hypothetical minimally competent examinee would rule out as incorrect. The reciprocal of the remaining number of options becomes each item's "Nedelsky rating"—that is, the probability that the F–D student would answer the item correctly. For example, on a 5-option item for which examinees would be expected to rule out two of the options as incorrect, the Nedelsky rating would be 1/(3 remaining options) = .33. The sum of the item ratings—or some adjustment to the sum—is used as a passing score. For example, a 50-item test consisting entirely of items with Nedelsky ratings of .33 would yield a recommended passing score of 16.5. Where the recommended passing score is not a whole number, it seems advisable to round the passing score up to the nearest whole number on the scale used to report results. In this case, only examinees who have attained a 17 (or greater) on the raw score scale can be said to have met or exceeded the passing mark of 16.5.

Serious limitations of the Nedelsky method have been described in the literature. For example, the method can only be used with the multiple-choice format. Other technical flaws have been noted, such as that the scale of Nedelsky values does not permit probabilities between .50 and 1.00 (Berk, 1984). Shepard (1980) has hypothesized that this is a reason that the Nedelsky method often results in standards that are lower than those obtained using other methods; judges tend not to assign probabilities of 1.0 (that is, to assert that *all* examinees will answer an item correctly).

Ebel's method. The methodology proposed by Ebel (1972) also requires participants to make judgments about test items. To implement the Ebel method, participants provide estimates of the difficulty of individual test items, judgments about the relevance of test content areas, and predictions about examinees' expected success on combinations of the difficulty and relevance dimensions. Commonly, participants are asked to categorize items according to three difficulty levels (easy, medium, hard) and four relevance levels (essential, important, acceptable, questionable). Participants then make judgments about how minimally proficient examinees will perform on the test, usually in the form of expected percentage correct for each difficulty-by-relevance combination.

Suppose, for example, that five participants provided the judgments shown in Table 1 for a 100-item test. Using the illustrated combination of judgments about difficulty, relevance,

Table 1
Illustration of Ebel Standard-Setting Method

Item category	Judged required mastery (A)	Number of items judged to belong in category (B)	A × B
Essential			
Easy	100%	94	9400
Medium	100%	0	0
Hard	100%	0	0
Important			
Easy	90%	106	9540
Medium	70%	153	10,710
Hard	50%	0	0
Acceptable			
Easy	80%	24	1920
Medium	60%	49	2940
Hard	40%	52	2080
Questionable			
Easy	70%	4	280
Medium	50%	11	550
Hard	30%	7	210
	Totals	500	37,630

Note. Adapted from Ebel (1972).

and expected success shown, the Ebel method would yield a recommended passing percentage of 37630 ÷ 500 = 75.26%—or 76 items correct.

One advantage of the Ebel method is that it can be used with item formats other than multiple choice. However, the method has also received criticism. For example, the method reveals inadequacies in the test construction process (e.g., Why should *any* items judged to be of questionable relevance be included on an examination?). It requires judgments that may not be necessary (e.g., empirical item difficulty values are often available). And, Berk (1984) has hypothesized that it may be hard for participants to keep the two dimensions of difficulty and criticality distinct, possibly because these dimensions are often highly correlated.

Angoff's method. Angoff's (1971) method, like the other item-based procedures, requires standard-setting participants to review test items and to provide estimation of the proportion of a subpopulation of examinees who would answer the items correctly. Angoff suggested that:

> A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical "minimally acceptable person" in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the "minimally acceptable person." (Angoff, 1971, pp. 514–515)

In practice, a footnoted variation to the procedure Angoff originally proposed has dominated applications of the method:

> A slight variation of this procedure is to ask each judge to state the *probability* that the "minimally acceptable person" would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities would then represent the minimally acceptable score. (Angoff, 1971, p. 515).

The Angoff method has become the most rigorously researched and widely used of the item-based procedures. In most instances, the procedure is modified to facilitate less variable estimations. For example, the so-called "modified Angoff" approaches often include two or more rounds of ratings. Such modifications—incorporated in many other methods besides the Angoff approaches—are often desirable because they provide an opportunity for participants to see how their ratings compare with other participants' ratings before generating final ratings.

It is also frequently recommended that participants be provided with normative data in one or more of the rounds of ratings. In the Angoff approach, this usually takes the form of actual item difficulty indices. This step is desirable as a means of promoting reasonable conceptualizations of anticipated examinee performance (although some standard-setting specialists have argued that such normative data degrade the criterion-referenced nature of the judgments participants are asked to make).

Table 2 shows a matrix of ratings to 10 items by 13 judges in two rounds of ratings. In this case, participants were instructed to imagine a group of 100 minimally competent examinees and to estimate the number who would answer a given item correctly. To make the task easier, participants were given a form on which to record their estimates. (In this case, the forms permitted estimates in multiples of 10 only, though this is not a requirement in Angoff procedures.) The upper and lower values in each cell are the first and second round ratings, respectively. The means for each judge and each item are also presented for each round. These values reveal that, in Round 2, Judge 10 produced the most lenient ratings (M = 63.0) and that Item 1 was judged to be the easiest (M = 88.5).

Derivation of a recommended passing score using the Angoff method is accomplished by averaging either the judge or item

Table 2
Illustration of Angoff Standard-Setting Method

	Judge													
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	Mean
1	90	90	100	100	100	90	90	90	90	60	90	100	90	90.8
	80	90	90	100	90	90	100	90	80	70	90	90	90	88.5
2	80	90	90	40	100	80	100	70	80	90	100	70	80	82.3
	80	70	90	60	100	80	90	80	70	80	80	80	90	80.8
3	90	70	80	80	100	60	80	80	80	60	50	90	80	76.9
	90	80	90	70	80	60	70	80	80	60	60	90	70	75.4
4	70	60	70	80	90	80	80	70	70	60	50	90	90	73.9
	70	70	60	70	80	80	70	70	70	70	70	80	80	72.3
5	90	60	90	40	80	60	80	70	60	60	90	70	80	71.5
	80	70	90	60	80	60	70	70	70	70	80	70	70	72.3
6	60	60	80	60	70	70	80	80	60	50	70	80	90	70.0
	70	60	70	70	70	70	70	80	60	50	70	80	90	70.0
7	90	50	80	60	60	70	70	70	70	60	80	80	70	70.0
	80	60	80	70	60	70	60	80	80	50	80	70	80	70.8
8	80	50	70	80	40	90	70	70	60	60	70	70	80	68.5
	70	50	80	70	50	90	70	80	70	70	70	80	80	71.5
9	80	70	60	70	60	80	50	60	60	30	50	60	90	63.1
	90	70	70	70	60	80	60	70	70	60	60	70	80	70.0
10	60	80	50	60	70	90	70	60	30	40	40	50	70	59.2
	70	80	60	70	80	90	80	70	40	50	60	60	60	66.9
Mean	79.0	68.0	77.0	67.0	77.0	77.0	77.0	72.0	66.0	57.0	69.0	76.0	82.0	72.6
	78.0	70.0	78.0	71.0	75.0	77.0	74.0	77.0	69.0	63.0	72.0	77.0	79.0	73.8

Note. Adapted from Engelhard and Cramer (in press).

means; usually the calculations are based on the final round of ratings, under the assumption that the ratings converge toward consensus and become less variable round-to-round. Using the Round 2 ratings shown in Table 2, the recommended passing score would be 73.8% correct—or 8 of the 10 items on the test.

Research on the Angoff method has suggested that it provides easy-to-obtain and acceptable results in many situations. For example, Mills and Melican report that "the Angoff method appears to be the most widely used. The method is not difficult to explain and data collection and analysis are simpler than for other methods in this category" (1988, p. 272). Colton and Hecht compared the Angoff, Ebel, and Nedelsky method*ologies and* reported that "the Angoff technique and the Angoff consensus techniques are superior to the others" (1981, p. 15). Berk concluded that "the Angoff method appears to offer the best balance between technical adequacy and practicability" (1986, p. 147).

One advantage of the Angoff method is that it can be applied to a variety of situations, including constructed response formats. In these modifications, participants generate expected scores for minimally proficient examinees on whatever score scale is used. For example, Hambleton and Plake (1995) describe the use of an "extended Angoff procedure" to set standards on performance assessments.

The purported ease of implementation—indeed, the validity—of the Angoff method has, however, recently been challenged. For example, a report of the National Academy of Education studied implementation of a modified Angoff approach used to set standards for the National Assessment of Educational Progress (NAEP). The report provided some evidence related to the inability of standard-setting participants to form and maintain the kinds of conceptualizations required to implement item-based procedures, suggesting that abstractions, such as minimally competent or borderline candidate, may be impossible for participants to acquire or to adhere to once acquired. The report also criticized the Angoff method as not allowing participants to adequately form integrated conceptions of proficiency. The report concluded that the Angoff procedure was "fundamentally flawed" and recommended that "the use of the Angoff method or any other item-judgment method to set achievement levels be discontinued" (Shepard, Glaser, Linn, & Bohrnstedt, 1993, p. xxiv). To date, these hypotheses have not received much empirical attention, and it is likely that item judgment methods will continue to see widespread use in the near future.

Jaeger's method. Jaeger (1982) developed another item-based procedure, similar to that initially suggested by Angoff (1971). To implement this procedure, participants answer the following question for each item in the examination: "Should *every* examinee . . . be able to answer the test item correctly?" (Jaeger, 1989, p. 494). Like some modifications of the Angoff method, Jaeger's procedure requires iterations of data collection, with participants provided an opportunity to reconsider their initial judgments after receiving information about the judgments of other participants and about actual examinee performance (e.g., anticipated pass/fail rates).

One advantage of the Jaeger procedure is its explicit recognition of the fact that various constituencies have a stake in the results of the standard-setting process. The procedure requires sampling from each population with an informed, legitimate interest in the outcome. To compute the actual passing score, the median standard for each sample of participants is calculated, and Jaeger suggests using the lowest of these as the recommended standard.

Like the Nedelsky procedure, the Jaeger procedure has been criticized as not allowing participants to make probability choices other than 0 or 1 (Berk, 1986); it may also produce somewhat less reliable standards than other item-based approaches (Cross, Impara, Frary, & Jaeger, 1984). However, because the Jaeger method has been introduced more recently, it has received comparatively less scrutiny than the Angoff, Ebel, or Nedelsky approaches.

Examinee-Centered Models

In contrast to procedures that require participants in the standard-setting process to make judgments about test items, examinee-centered methods require participants to make direct judgments about the status of persons on the construct of interest (e.g., competent/not competent). To derive a passing score for the test, the judgments are combined with information about the performance of the same group of persons on an examination. The examinee-centered methods differ in how the information is combined to arrive at the passing score.

It is sometimes suggested that examinee-centered methods represent a more natural approach to setting standards. Making judgments about item content may be difficult for standard-setting participants because it is a more contrived task (Poggio, Glasnapp, & Eros, 1982); Livingston and Zieky (1989) have also suggested that the main advantage of examinee-based methods is that standard-setting participants are likely to be more accustomed to judging students' abilities as being adequate or inadequate for a particular purpose than they are with estimating probabilities. Another advantage is that actual performances of real people are judged, as opposed to eliciting estimates about the probable performance of a hypothetical group. In the following sections, the two most frequently cited examinee-based methods, the contrasting groups and the borderline group approaches, are described and illustrated.

Contrasting Groups method. The contrasting groups method was described under another name by Berk who suggested "an extension of the familiar known-groups validation procedure" (1976, p. 4). This method involved administration of an examination to two groups of students—those who were known to have received effective instruction covering the content to be tested and those who had not. The two distributions of test performance could be examined to find a point on the score scale that maximized the probability of correct decisions (i.e., identifying true masters and nonmasters) and minimized the probability of incorrect decisions (i.e., identifying false masters and nonmasters).

A variation of Berk's method involves asking participants, who have knowledge of both the examinee population and the required knowledge or skill level, to classify examinees as either competent or not competent. Livingston and Zieky (1982) recommend plotting the percentage of test takers at each score level who are judged to be competent. To derive a passing score they recommend that "one logical choice is the test score for which the 'smoothed' percent-qualified is exactly 50 percent" (p. 40). Another possibility is to select a point that minimizes the overall impact of errors of classification. The upper portion of Figure 1 illustrates a passing score obtained using the contrasting groups method, with the cutting score indicated as C_x.

One concern about the contrasting groups method is the validity and dependability of the criterion judgments. For example, judgments assigning examinees to "known" master or nonmaster groups are fallible. It is equally necessary to examine the adequacy of *these* classifications as it is to examine the psychometric characteristics of the predictor (i.e., the examination).

Also, the procedure for ultimately deriving a recommended standard described in this section is referred to as the contrasting groups graphing method. However, other methods using the same contrasting groups approach have been suggested, including the decision-making accuracy approach (Berk, 1976), base rates analysis (Peters, 1981), utility function analysis (Overall & Klett, 1972), and discriminant function analysis (Koffler, 1980). The relative advantages of these alternatives have not received much attention in the literature.

Borderline Group method. Zieky and Livingston (1977) proposed using a single group judged to be at the borderline separating competent from noncompetent performance. To implement the procedure, participants who are familiar both with examinees at this level and with the knowledge or skills to be tested identify a sample of members of this subpopulation. The median score of this sample can be used as a recommended standard. The lower portion of Figure 1 illustrates a passing score obtained using the borderline group method.

One advantage of the borderline group method is its intuitive nature. Like the contrasting groups method, it requires the kind of judgments that participants are likely used to making. On the other hand, it is often difficult to identify a borderline group of sufficient size. And, Jaeger has observed that participants possessing enough familiarity with the examinee group to make such judgments "are likely to be influenced by cognitive and noncognitive factors that fall outside the domain assessed by the test" (1989, p. 497). He suggests that participants' judgments are also likely to be adversely affected by errors of central tendency, placing examinees for whom they have insufficient information into the borderline group.

Compromise Models

Another family of standard-setting methods was introduced following initial attempts by Nedelsky (1954) and others to determine "absolute" passing standards. These models aspired to develop methods that would strike a compromise between purely norm-referenced (relative) approaches and absolute methods. The methods can be used to derive passing scores outright or to adjust standards obtained using other methods.

Compromise models have been suggested by Beuk (1984), deGruijter (1980), Grosse and Wright (1986), and Hofstee (1983). Overviews of these methods are provided in deGruijter (1985) and Mills and Melican (1988); however, little comparative work has been done to establish advantages and disadvantages of these models. The following sections describe two of the more commonly encountered methods, the Beuk and Hofstee approaches.

Beuk's method. As Beuk (1984) has observed, "setting standards . . . is only partly a psychometric problem" (p. 147); he suggests that standard-setting procedures take into account both the content requirements necessary for acquisition of a

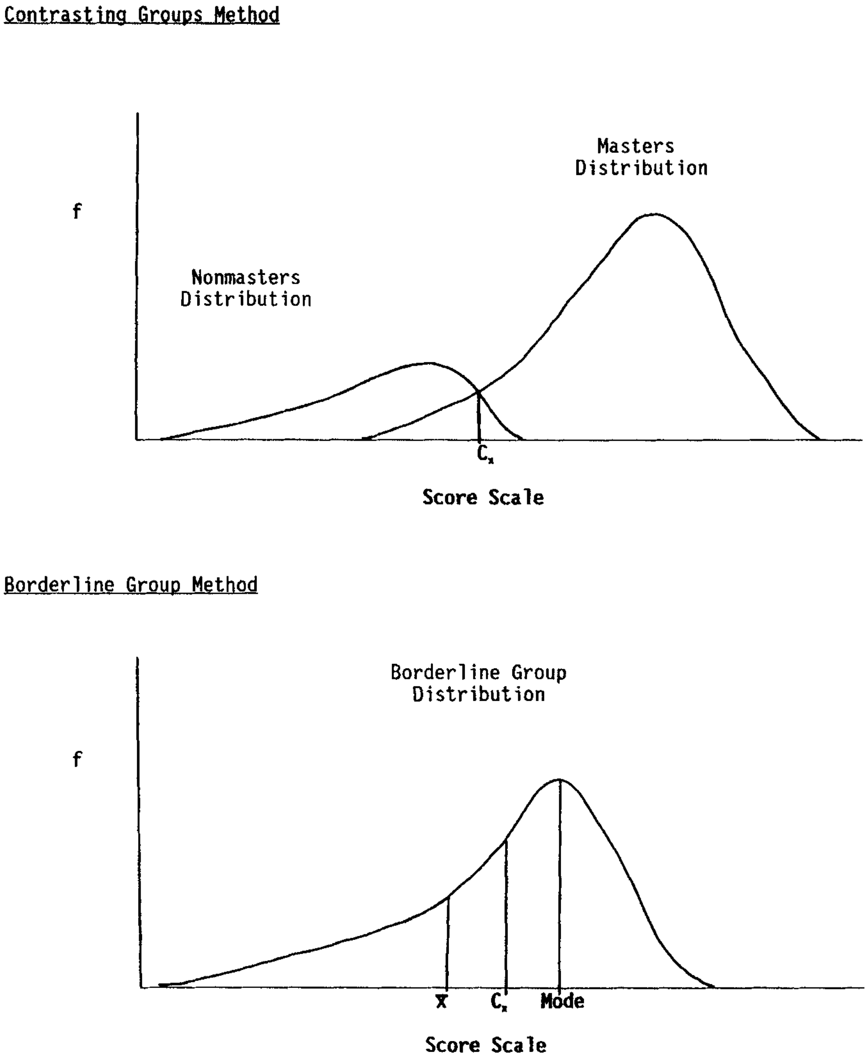

FIGURE 1. *Illustrations of examinee-based methods*

credential (i.e., absolute information) and comparative achievement of participating examinees (i.e., relative information).

To implement Beuk's (1984) method, each participant in the standard-setting procedure is asked to make two judgments: (a) the minimum level of knowledge required to pass an examination, expressed as a percentage of the total raw score on the test and (b) the passing rate expected, expressed as a percentage of the examinee population. When the examination has been administered, these expectations can be compared with reality. If the expectations differ from reality, a compromise between the two can be struck using the information provided by the participants' judgments.

The upper portion of Figure 2 provides a conceptual illustration of deriving a passing score using Beuk's method. In the figure, the intersection of the mean expected pass rate and the mean expected percentage correct (labeled Point A) is used as a reference point. A line with a slope equal to the ratio of the standard deviations of participants' judgments about expected knowledge levels and passing rates is passed through Point A and projected onto a curve showing the functional relationship between percentages of successful examinees and possible cutting scores. The point at which the line intersects the curve (labeled Point B) is used to derive the recommended passing percentage and the consequent passing rate. The recommended passing score can be obtained by multiplying the adjusted percent correct ($\bar{X}'$) by the number of items in the examination.

Hofstee's method. Like Beuk, Hofstee observed that important classification decisions are "based on two classes of premises, one political and the other cognitive" (1983, p. 109). Hofstee's method is also an attempt to strike a compromise between these competing perspectives. He originally proposed a model which would apply "to the situation in which a cutoff score on an achievement test is set for the first time. . . . [when] no agreed-upon prior or collateral information is available on the difficulty of the test, the quality of the course, or the amount of preparation by the students" (p. 117). In theory, the approach could also be used when such information is available.

The Hofstee (1983) approach is implemented by asking each standard-setting participant to respond to four questions: (a) What is the lowest cutoff score that would be acceptable, even if every student attained that score on the first testing? (b) What is the lowest acceptable cutoff score, even if *no* student attained that score on the first testing? (c) What is the maximum tolerable failure rate? (d) What is the minimum acceptable failure rate? These means across judges of these values are referred to, respectively, as $k_{\min}$, $k_{\max}$, $f_{\max}$, and $f_{\min}$.

To derive a cutting score, the points ($f_{\min}$, $k_{\max}$) and ($f_{\max}$, $k_{\min}$) are used to plot a line which, like the Beuk (1984) method, is projected onto the distribution of observed test

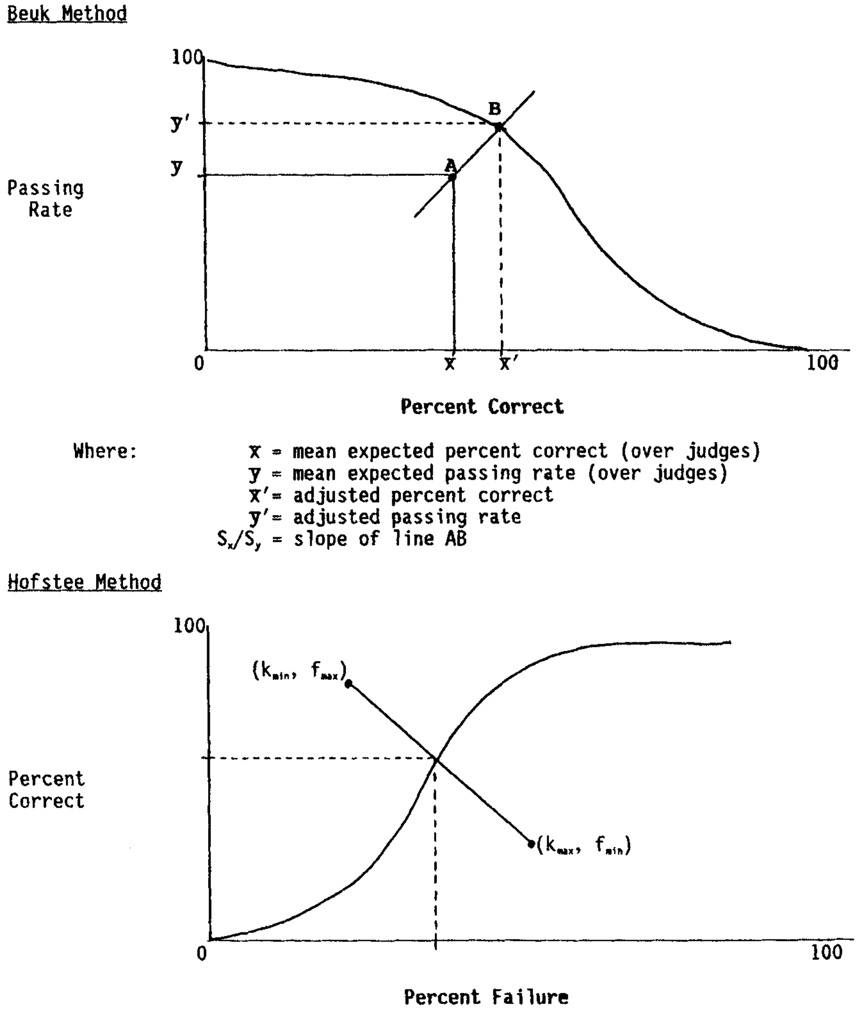

FIGURE 2. *Illustration of compromise methods*

scores. The lower portion of Figure 2 illustrates the locations of k_{min}, k_{max}, f_{max}, and f_{min} and shows the projection of the resulting line onto a curve which shows the functional relationship between percentages of failing examinees (on the abscissa) and the percentage of correct responses on a test (on the ordinate).[1] The passing percentage is found by following the dashed line to the ordinate; the corresponding failing rate is found by following the dashed line to the abscissa.

The Hofstee method may be used with any item format, may be instituted singly, or utilized as supplementary information when another method is used. As Jaeger (1989) has observed, however, one disconcerting possibility when using the Hofstee method is that the line may not intersect the test score distribution, resulting in a failure to identify a solution to the passing score problem.

Other Issues

In the actual practice of setting standards, the devil is in the details. A few of the other important questions that must be addressed are presented in this section. For example, the use of the lowest median as the passing score in the Jaeger (1982) method described earlier tacitly assumes certain beliefs about the relative costs of false positive and false negative decisions. In any passing score study, it would be desirable to discuss these values and potential costs as a part of participants' training.

The extent to which participants should interact in making their judgments is also a concern. Fitzpatrick (1989) reviewed literature related to social influences in standard setting. She suggests the following conclusions derived from the literature that may help frame the way interaction is incorporated into standard setting studies:

1. When participants are initially disposed to favor one position over another, discussion of the issue or exposure to the other position will tend to polarize their opinions, with subjective judgments more susceptible to polarization than objective judgments and the polarizing effect of discussion greater than that of exposure;
2. Exposure to an extreme group norm or mean opinion position induces more polarization than simple exposure to a distribution of opinion positions; and
3. Several strategies are known to mediate polarization, including private recording of judgments, carefully structuring the discussion among participants, and reducing the subjectivity of the judgments they are asked to make.

Another concern is the optimal number of iterations in a standard-setting process. Further research is required to understand when the limit of variation reduction is approached. Perhaps the least well researched concern is training that should be provided to participants—particularly, how much time should be devoted to training, what training methods are most effective, and how to evaluate the effectiveness of the training. Cizek (1996) has suggested that a potentially fruitful line of research lies in combining the needs of standard setters with the knowledge base in instructional design.

Standard Setting and New Modes of Assessment

A new way of classifying standard setting models has been proposed by Kane, who suggests that they be viewed as representing the presumed cognitive models of competence underlying the testing program. He suggests *holistic models* "which assume that achievement or skill is highly integrated" and *analytic models* "which assume that achievement can be assessed using relatively small parts or samples of performance" (1994b, pp. 4–5).

Clearly, the future of standard setting is fixed on a course oriented toward more adequately reflecting the nature of knowledge and skill acquisition in a given domain. Many of the methods described earlier were developed for applications with the multiple-choice item format and with the implicit assumption of an analytic model. Until recently, measurement research has addressed standard setting primarily in the context of written examinations, with less attention devoted to the problems of setting standards for actual demonstrations or performances. Some of the existing methods have been applied to other formats, but with mixed results. The recent increased emphasis on complex, performance-based assessments seems to point toward the need for development of holistic models.

Frameworks and methodological options for setting standards for new modes of assessment are currently being developed and described (see, e.g., Jaeger, 1994; Poggio & Glasnapp, 1994; Putnam, Pence, & Jaeger, 1995). One example of a new method called *policy capturing* has been proposed by Jaeger and his colleagues, who studied its application in the context of a teacher certification assessment. The assessment utilizes several dimensions of performance involved in early adolescence English language arts instruction and provides standard-setting participants with a framework for developing acceptable and unacceptable profiles of performance across the dimensions. Because the characteristics assessed are assumed to be complex performances, and because participants possess holistic notions of acceptable performance, the standard-setting method has been conceived to reflect those assumptions.

An initial classification and explication of methods for setting standards on examinations using constructed response formats is provided by Faggen (1994). Faggen's work reflects the fact that much standard setting for constructed response examinations has been concerned with establishing performance levels in the area of writing assessment. Faggen describes four procedures: (a) benchmark; (b) item-level pass/fail; (c) item-level passing score; and (d) test-level pass/fail.

The benchmark method requires standard-setting participants to be familiar with examinee performances (e.g., essays) that clearly portray points along the scoring continuum; these are called "benchmark" performances. Then, individual participants identify scale values that indicate minimally competent performance; the mean of individual participants' passing scores is used as the recommended standard. The process may or may not include an iterative process of providing information about the judgments of other participants and additional rounds of ratings.

The item-level pass/fail method involves reviewing a representative sample of examinee responses to each item or task in an examination. Participants then judge, for each item or task, whether the performance should be assigned to a pass or fail category. For example, a participant who reviewed a sample of 15 examinee responses to 4 constructed response items would provide a total of 60 pass/fail judgments. Like the benchmark method, the item-level pass/fail method might include an iterative component. A recommended passing standard is derived by estimating "for each possible score that a response could receive . . . the probability that a randomly selected response with that score would be judged 'Pass' by a randomly selected member of the standard-setting panel" (Faggen, 1994, p. 5). Faggen suggests the use of logistic regression to obtain, for each question, the raw score associated with a .50 probability of being judged a "pass," a suggestion similar to the contrasting groups approach.

Faggen also describes an item-level passing score method, which resembles an adaptation of the Angoff (1971) method for use with constructed response formats. The method is also similar to the item-level pass/fail method, except that, instead of asking participants to specify a point on the score scale that represents acceptable performance, they are asked to estimate the average scale value that would be attained by a group of minimally competent examinees.

The fourth method described by Faggen is the test-level pass/fail method. As with other methods, implementing this procedure begins with training to acquaint qualified participants with the examination items or tasks, the scoring guidelines or rubrics, and benchmark responses. This method uses an iterative process of providing holistic pass/fail judgments on samples of test or whole-booklet performance. Faggen describes implementation of the method on a writing examination in which "panel members examine three sets of examinee test booklets . . . with two papers at each of three different score points, one high, one low, and the third in the middle at a value likely to be in the passing score range" (1994, p. 8). Subsequently, participants are asked to provide overall pass/fail judgments for a set of papers with the same score, lying within their previous pass/fail range. The recommended passing score is obtained by taking the mean of the each participant's individual standard at the end of the iterative process.

Standards and Validity Evidence

Standards, like tests themselves, are not stamped with an imprimatur of "valid." Validity in standard setting does not exist outside of the value systems that define what are desirable outcomes: What is considered "reasonable" or "appropriate" ultimately depends on individual values. However, the framework for assessing validity provided by Messick (1989) suggests that the validity of standards can be evaluated by an on-going process of gathering and evaluating evidence that bears on the question of whether the inferences implied by application of a cutting score are warranted. Standard setting represents the point in the assessment process that yields critical classifications; these classifications ultimately result in inferences about competence, proficiency, or other important characteristics.

Kane has expressed the intimate relationship between validity and standard setting, noting that "validation . . . consists of a demonstration that the proposed passing score can be interpreted as representing an appropriate performance standard (1994a, p. 426). He further suggests that validation be accomplished by means of constructing plausible "interpretive arguments" that link test scores to score-based inferences or decisions (1992, p. 527).

Accordingly, standard-setting procedures must provide support for the validity argument. They should be designed to address the question of valid inferences and should yield evidence bearing on the accuracy of those inferences. One necessary, though not sufficient, aspect of the evidentiary trial is the proper following of a prescribed, rational system of procedures designed to synthesize inherently judgmental decisions. Among other potential sources of validity evidence are:

- the foundation of the assessment itself (e.g., foundation in task/job analysis, content validity evidence, etc.);

- evidence of clear definitions of key constructs used by participants in the standard-setting procedure (e.g., *minimal competence, proficient, borderline*);
- documentation regarding how the participants in the standard-setting procedure were sampled, selected, and trained;
- evidence regarding the quality of the materials used in the standard setting and the careful implementation of the standard-setting method;
- evidence that participants in the standard setting understood the method and applied it correctly; and
- evidence from external sources that the standard is reasonable and appropriate.

Conclusion

Theory and practice in standard setting have steadily developed, and measurement specialists have provided rational, systematic solutions to the practical testing problem of setting performance standards. Much additional work needs to be done, however. For example, research is continuing to address the factors that influence standard-setting participants (see, e.g., Fitzpatrick, 1989; Plake, Impara, & Potenza, 1994; Smith & Smith, 1988).

The practice of standard setting clearly involves judgment; it represents the nexus of research design, training, statistics, values, and policy considerations. As Shepard has observed, "All standard-setting is judgmental. Our empirical methods may facilitate judgment making, but they cannot be used to ferret out standards as if they existed independently of human opinions and values" (Shepard, 1979, p. 62). And, the exercise of judgment regarding minimally acceptable performance on complex cognitive tasks is surely difficult.

Measurement specialists can assist in the process of gathering and synthesizing judgment by enhancing the dependability and validity of standard setting; traditionally, these efforts have been directed toward reducing within-participant inconsistency and between-participant variability. As the move toward increasing reliance on complex, performance-based tasks continues, measurement specialists can also contribute by developing and refining procedures for setting standards on these assessments and by developing standard-setting models that are closely aligned with the cognitive models of competence inherent in the performances.

Notes

The author is grateful for support for this work provided by The University of Toledo College of Education and Allied Professions.

[1] Hofstee actually recommended that, for multiple-choice tests, the ordinate represent the percentage of right answers corrected for guessing.

Self-Test

1. Consider the four concerns that Linn (1994) suggests that standard setting can address. Is it possible for a single standard-setting process to address more than one concern?

2a. One reason for including iterations in the generation of item ratings is to reduce variability in the ratings. In the ratings shown in Table 2, did the Round 2 ratings become less variable? How do you know?

2b. What would the passing score have been if only Round 1 ratings had been collected?

3. Suppose that most participants produced lower second-round ratings (e.g., 5, 10, 15) when providing Angoff estimates for 4-option multiple-choice items. How will these ratings affect the recommended passing score compared to the passing score from the first round? What deficiency in training do these ratings reveal?

4. Suppose that physicians are videotaped while interacting with patients and that the performances are rated by a group of experts. What standard-setting procedure could be used to establish a cutoff for acceptable and deficient interaction skills?

5. Suppose that a standard-setting panel using the Ebel method arrives at a recommended passing score that seems unreasonable (e.g., too many examinees will fail if the cut score is used). What are some options that could be recommended to address this situation?

6. Suppose that a 5-point scale is established to rate student writing. In an assessment, two samples of writing (narrative and expository) from each student are collected. What method(s) could be used to establish a passing score on the writing assessment, and how could the reliability of the method be evaluated?

7. What should be included in a report of a standard-setting study?

Self-Test Answers

1. Yes. As Linn notes, the purposes are not mutually exclusive. For example, a competency test for high school graduation may be meant to spur achievement on the part of students (exhortation), to communicate high expectations or aspirations for students (exemplification), to drive curricular change and alter instructional practice (accountability), as well as to meet its stated purpose of certifying student competence.

2a. Yes. The Round 2 ratings became less variable as evidenced by the change in the variability of ratings from Round 1 (S=6.90) to Round 2 (S=4.65). However, whether or not this particular reduction in variability is desirable is another matter. For example, the reduction might have been attributable to domination of group interaction by a single participant. Careful monitoring and structuring of such interactions would provide insights into the observed reduction in variability.

2b. The Round 1 recommended passing score would have been 72.6% correct.

3. The lower ratings would result in a lower recommended passing score compared to the initial ratings. In this case, the low ratings may reveal that the participants were not informed about the possibility of guessing a correct response or that they failed to take this information into account. (Recall that the expected percentage of examinees answering an item correctly when only blind guessing is involved would be 25%.)

4. There are many alternatives. For example, if judgments about the competence of the physicians are also available, the contrasting groups or borderline group procedure could be used. Faggen's test-level pass/fail method could also be used.

5. Situations like the one described may be more common than we might wish. To address the problem, one could consider: using another method, such as Jaeger's (1982), which explicitly incorporates information about how pass/fail rates are affected by participants' judgments; calculating the passing score without the ratings of participants who are judged to have produced "unacceptable" ratings, based on examination of statistical indices (see, e.g., Harnish & Linn, 1981; Engelhard & Cramer, in press); using an adjustment to the passing score, such as the Beuk or Hofstee methods; or constituting another panel, incorporating revised—hopefully improved—training procedures.

6. Again, many methods would be possible. The methods described by Faggen (1994) were developed specifically for purposes such as the one described. Perhaps the

most interesting aspect of this situation is the ability to apply generalizability theory to examine the dependability of the ratings and the relative contributions of raters and prompt type to variability in the ratings (see, e.g., Brennan & Lockwood, 1980).

7. Reporting on standard setting should be as full, open, and accessible as the situation permits. Wherever relevant, documentation should include reference to applicable standards in the *Standards for Educational and Psychological Testing* (AERA/APA/ NCME, 1985). Also, although no formal standards for standard setting have been adopted, Cizek (1996) has suggested guidelines for documenting standard setting related to the purposes, methods, procedures, and analyses used. These guidelines include:

Purpose	• Define the purpose for setting standards • Define relevant constructs
Method	• Connect the purpose and method of setting standards • Connect the characteristics assessed and method • Describe the standard-setting method selected
Procedures	• Describe procedures as implemented • Describe adjustment procedures, if used
Technical and Procedural Analysis	• Describe the participant group and method of selection • Present evidence of participants' task comprehension • Document appropriate use of information by participants • Report magnitude of error

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

Berk, R. A. (1976). Determination of optional [sic] cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45*(2), 4–9.

Berk, R. A. (1984). *A guide to criterion-referenced test construction.* Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137–172.

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

Block, J. H. (1978). Standards and criteria: A response. *Journal of Educational Measurement, 15,* 291–295.

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

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30,* 93–106.

Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practice, 15*(1), 13–21, 12.

Colton, D. A., & Hecht, J. T. (1981, April). *A preliminary report of three techniques for setting minimum passing scores.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Los Angeles.

Cone, J. D., & Foster, S. L. (1991). Training in measurement: Always the bridesmaid. *American Psychologist, 46,* 653–654.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21,* 113–129.

deGruijter, D. N. M. (1980, June). *Accounting for uncertainty in performance standards.* Paper presented at the International Symposium on Educational Testing, Antwerp.

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

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

Engelhard, G., & Cramer, S. E. (in press). Using Rasch measurement to evaluate the ratings of standard-setting judges. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice.* Norwood, NJ: Ablex.

Faggen, J. (1994). *Setting standards for constructed response tests: An overview.* Princeton, NJ: Educational Testing Service.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59,* 315–328.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237–261.

Grosse, M. E., & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation & the Health Professions, 9*(3), 267–285.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41–55.

Harnish, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133–146.

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

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461–475.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan.

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

Jaeger, R. M. (1994, April). *Setting performance standards through two-stage judgmental policy capturing.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

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

Kane, M. T. (1994b, October). *Examinee-centered versus task-centered standard setting.* Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement, 17,* 167–178.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education.* Cambridge, MA: Harvard University Press.

Linn, R. L. (1994, October). *The likely impact of performance standards as a function of uses: From rhetoric to sanctions.* Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores.* Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*(2), 121–141.

Madaus, G. F. (1992). A national testing system: Manna from above? Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research, 45,* 133–158.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Mills, C. N., & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. *Applied Measurement in Education, 1,* 261–275.

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

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

Overall, J. E., & Klett, C. J. (1972). *Applied multivariate analysis.* New York: McGraw Hill.

Peters, E. (1981). *Basic skills improvement policy implementation guide #3: Standards-setting manual.* Boston: Massachusetts State Department of Education.

Plake, B. S., Impara, J. C., & Potenza, M. T. (1994). Content specificity of expert judgments in a standard-setting study. *Journal of Educational Measurement, 31,* 339–347.

Poggio, J. P., & Glasnapp, D. R. (1994, April). *A method for setting multilevel performance standards on objective or constructed response tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Poggio, J. P, Glasnapp, D. R., & Eros, D. S. (1982, March). *An evaluation of contrasting groups methods for setting standards.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement, 15,* 297–300.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 57–83.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10*(2), 11–14.

Shepard, L. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based education* (pp. 59–71). Washington, DC: National Council on Measurement in Education.

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*(3), 447–467.

Shepard, L. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169–198). Baltimore, MD: Johns Hopkins University Press.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement.* Stanford, CA: National Academy of Education.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using the Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25,* 259–274.

Zieky, M. J. (1994, October). *A historical perspective on setting standards.* Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

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

Annotated Bibliography

Educational Measurement: Issues and Practice. (1991). *10*(2). This special issue has several articles which address various aspects of standard setting, from selection of standard-setting participants to factors influencing participants' judgments and how standard-setting data can be combined to arrive at recommended passing scores.

Educational Measurement: Issues and Practice. (1994). *13*(4). This special issue recounts some of the history of the development of criterion-referenced testing, its applications, and prospects and problems for the future.

Faggen, J. (1994). *Setting standards for constructed response tests: An overview.* Princeton, NJ: Educational Testing Service. This brief research memorandum provides a succinct, practitioner-oriented categorization and explication of four strategies for setting standards on tests using constructed response formats (e.g., essays, calculations, performances). The style is how-to but lacks some of the attention to difficult issues presented in the primary practitioner reference by Livingston and Zieky (1982).

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan. This chapter is the beginning required reading for anyone interested in standard setting. Jaeger provides an excellent overview of methods, issues, and concerns for the future.

Journal of Educational Measurement. (1978). *15*(4). This special issue is a foundational work for anyone interested in understanding the debate about the nature of setting performance standards. It includes the classic essay by Glass on "Standards and Criteria" as well as several other thought provoking responses and perspectives.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores.* Princeton, NJ: Educational Testing Service. The original and still the only practitioner-oriented handbook for conducting passing score studies. Many improvements in recommended practice have occurred since the publication of *Passing Scores,* but most of the information it contains is still valuable. This handbook goes beyond simple how-to and provides foreshadowing of important issues that must be addressed in standard setting, such as qualifications of participants, sampling, options for data analysis, etc.

Teaching Aids Are Available

A set of teaching aids, designed by Gregory J. Cizek to complement his ITEMS module of "Setting Passing Scores" is available at cost from NCME. These teaching aids consist of brief additional text and 19 overhead masters. As long as they are available, they can be obtained by sending $5.00 to: **Teaching Aids, ITEMS Module #18, NCME, 1230 17th St., NW, Washington, DC 20036.**

Call for Nominations for the NCME Outstanding Dissertation Award

The National Council on Measurement in Education announces its first annual competition for the best doctoral dissertation in the field of educational measurement. Nominations will be accepted for dissertations completed between July 1, 1994, and June 30, 1996. Membership in NCME is not required; however, the student's advisor must be a member of NCME.

The criteria used by the Dissertation Award Committee to judge the nominated dissertations will include: significance of the contribution to research in the field, quality of the literature review, appropriateness of the design and methodology, technical quality of the research, quality and clarity of the writing.

After reviewing all nominations, the Dissertation Award Committee will select the top five dissertations. Copies of the complete dissertation will be requested in order to make the final selection of the most outstanding dissertation. The Committee may elect not to give an award. The award will be presented at the following NCME Annual Meeting in 1997. The winner will receive a $500 monetary award and a plaque. In addition, the advisor or committee chair of the winning dissertation will receive a congratulatory letter. An announcement of the award recipient will be published in the NCME newsletter. An honorable mention may also be awarded; its recipient will be recognized with a certificate.

To nominate a dissertation, the following items must be submitted to the Chair of the Dissertation Award Committee by July 31, 1996: (a) a summary of the dissertation research (up to 10 pages), including a brief overview of the literature, major research question, design and methodology of the study, results, and discussion —the reference list should be included as well—(b) a table of contents (including a list of tables and figures); (c) a letter of nomination from the student's advisor; (d) a statement from the graduate school to confirm the date of completion and acceptance.

Submit materials to: Dr. Jeri Benson, Chair of the NCME Dissertation Award Committee, Department of Educational Psychology, 325 Aderhold, University of Georgia, Athens, GA 30602. Phone: 706-542-3818; e-mail: cmsjeri@uga.cc.uga.edu