דוח המרכז

228

Standard-Setting in Testing: A Survey of the Literature

Aliza Berger

December 1996

מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
מיסודן של האוניברסיטאות בישראל
ת"ד 26015 ירושלים 91260, טל' 02-6759555

TABLE OF CONTENTS

INTRODUCTION

Perspectives and Definitions

In educational measurement, "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 [discontinuous] grades (e.g., A,B,C,D,F)[1] 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).[2]

"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.

"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 (Jaeger, 1989). "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 (1978), who held that attempts to set standards were 'either blatantly arbitrary or derive[d] from a set of arbitrary premises' . He called the decision making process of standard setting 'judgmental, capricious, and essentially unexamined'. 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. Shepard (1984) observed, 'the standard we are

---

[1] On this scale, A denotes excellent performance and F denotes failure.

[2] Material in quotation marks in the Introduction is taken from Cizek, 1996b, unless otherwise noted.

groping to express is a psychological construct in the judges' minds.'

"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'. 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 can be made.

"A second, *conceptual* definition is provided by Kane (1994), 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.' "

Guidelines for Standard Setting

"The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1985) 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 at critical score levels, especially at or near the cut score (Standard 2.10)."

"However, it is immediately discernable that the *Standards* does not present an integrated, unified perspective on standard setting that would parallel its treatment of other topics (such as validity, scaling, etc.). The six references to standard setting are sprinkled throughout the *Standards* and seem variable and disjointed, often treating standard setting only in the context of other testing issues (e.g., validity, testing for certification, etc.). Further, the *Standards* is silent on many critical aspects of standard setting" (Cizek, 1996a).

Cizek (1996a) offers guidelines for documenting the standard-setting process and, "by extension [addresses] the issues and concerns that should *precede* the actual conduct of a standard-setting procedure." Cizek draws an analogy between the standard-setting process and a research study; "the documentation of a standard-setting study should follow the same principles that guide any research report." The format of a research report (Purpose, Method, Procedures, Analysis) was followed in Cizek's guidelines, presented in Table 1.

Table 1
Recommended Guidelines for Standard Setting (Cizek, 1996a)

I. Purpose
    A. Define the purpose for setting standards
    B. Define relevant constructs
II. Method
    A. Connect purpose and method of setting standards
    B. Connect characteristics assessed [e.g. item format] and method
    C. Describe the standard-setting method selected
III. Procedures
    A. Describe procedures as implemented
    B. Describe adjustment procedures
        1. Adjustments to participants
        2. Adjustments to judgments
        3. Adjustments to passing scores
IV. Technical and procedural analysis
    A. Describe the participant group [judges] and method of selection
    B. Present evidence of [judges'] task comprehension
    C. Document appropriate use of information [about the test] by participants
    D. Report magnitude of error

Structure of the Report

In this report, general material on standard-setting will be presented, following in part the topics in Cizek's (1996a) "Guidelines for Standard-Setting". This background material will be followed by analyses of standard-setting in two practical applications. In Chapter 1, a description and critique of the most common standard-setting methods will be presented. Chapter 2 will cover the material included under Cizek's "Procedures" and "Technical and Procedural Analysis". In Chapter 3, validity issues will be considered. Chapter 4 will consist of a report and discussion of standard-setting procedures used on NAEP. In Chapter 5, an analysis of medical board certification tests will be presented. The summary will include recommendations for standard-setting on NAEP-like and medical certification tests based on the information gathered for this report. One issue that has recently received increased attention in the literature, standard setting for performance assessments (e.g. a special issue of *Applied Measurement in Education*, 1995), will not be covered in detail because it is less central to the applications for which this report was prepared.

CHAPTER 1: STANDARD-SETTING METHODS

The intermediate product of a standard-setting process, before the final passing score is set, is the standard-setting data. The quality of these data is probably the most important measurement consideration in the standard-setting process (Cizek, 1996b). In this chapter, the most commonly used *standard-setting methods,* i.e. methods for collecting the data and setting the passing score, will be described. After each method is described, its documented advantages and disadvantages will be listed. Following these descriptions, an overall evaluation will be presented. To focus the reader's efforts in this chapter, it is noted at the outset that the major methods referred to in later chapters are the commonly-used Angoff method and the new Bookmark method.

Early standard setting often utilized norm-referenced, *relative* methods. "For example...a credentialing examination might establish a passing mark at one standard deviation below the mean score for the group"[3]. "By the 1970s, with the proliferation of criterion-referenced testing, relative methods for setting standards were displaced by so-called *absolute* methods....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 the cutting score. State models "have not found wide applicability in competency testing programs" (Jaeger, 1989).

Continuum models are the most widely used in educational testing today. A further categorization of continuum models was into *judgmental* or *empirical* models (Hambleton & Eignor, 1980). Judgmental models involved judgments about test items and empirical models involved judgments about examinees. However, because both models employ judgment, Jaeger (1989) termed them instead *test-centered* and *examinee-centered* models.

Test-centered Models

"Just as criterion-referenced testing has proliferated, so too have the number of standard-setting methods available for these instruments. By [1986], Berk had catalogued 38 alternatives in a "consumer's guide to setting performance standards (1986)." Many of these alternatives are variations on basic methods. The four most

---

[3] Material in quotation marks in this chapter is taken from Cizek, 1996b, unless otherwise noted.

commonly used methods, and one new method, are presented in the following sections. Table 2 lists the distinguishing feature(s) of each method.

Table 2

Distinguishing features of test-centered standard-setting methods

Method	Feature
Nedelsky (1954)	Takes into account attractiveness of distractors; can only be used for multiple-choice tests
Ebel (1972)	Takes into account relevance of items
Angoff (1971)	Most often used and researched; many variants exist
Jaeger (1982)	Explicitly recognizes various stakeholder constituencies
Bookmark Approach (1996)	Items are presented to judges in rank order of difficulty; cutpoints set as "bookmarks" along the ordering

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"[4] 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 should therefore 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).

"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. [This is determined according to the judge's opinion.] The reciprocal of the remaining number of options

[4] F denotes a failing grade and D denotes a just passing grade on a scale of A,B,C,D,F.

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."

Advantages of the Nedelsky method.

1) It is easy to explain and computationally simple (Mills & Melican, 1988).

2) It explicitly takes into account one very real aspect of item difficulty: attractiveness of distractors (Gross, 1985).

Limitations of the Nedelsky method.

1) "The method can only be used with the multiple-choice format."

2) "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)."

3) "The estimated item difficulties cannot vary along the full range of difficulty, but are limited to discrete points as defined by the reciprocal of the number of options not eliminated by the judge (Brennan & Lockwood, 1980).

4) "The assumption of random guessing among options not eliminated by the judges has not been shown to hold (Melican, Mills & Plake, 1987).

5) "Judges find the Nedelsky method difficult to use for some types of items (e.g., items that are negatively worded).

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 3 for a 100-item test. Using the illustrated combination of judgments about difficulty, relevance, and expected success shown, the Ebel method would yield a recommended passing percentage of 37630 / 500 = 75.26% or 76 items correct."

Since predicting test scores is arguably a more familiar task for judges than predicting performance on single items, it was assumed that judges are more accurate with the Ebel method than with others. However, actually, there is no evidence that judges are more accurate with the Ebel method than with other methods.

Advantage of the Ebel method.

"It allows each item to be evaluated not only on its difficulty, but also on its relevance" (Mills & Melican, 1988). (But see Criticism 1 below).

Table 3
Illustration of Ebel Standard-Setting Method

Item category	Judged required mastery (A)	Number of items judged to belong in category (B)	A x 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

(Cizek, 1996b)

Criticisms of the Ebel method.

1) "The [basic assumption of 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?).

2) It may be hard for judges to keep distinct the two dimensions of difficulty and relevance (Shepard, 1984).

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 [average] 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)." A compromise, as in the worked-out example here, is to provide the normative information only at a later iteration in the rating process. [See more on use of empirical data by judges in Chapter 2.]

"Table 4 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 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 4, the recommended passing score would be 73.8% correct–or 8 of the 10 items on the test."

Examples of other kinds of information provided to judges are: extent of between-participant agreement (e.g., interjudge consensus data); extent of within-participant stability (e.g., intrajudge consistency data); information on the impact of participants' judgments (e.g., projected pass/fail rates) In general, "empirical work supports the notion that providing such information helps to make judgments more accurate, consistent, and realistic" (Jaeger, 1982; Busch & Jaeger, 1990; Livingston & Zieky, 1982; Shepard 1980).

Item difficulties can be provided for the overall group or for the borderline group. However, providing statistics for the borderline group raises the issue that although the judges are supposed to be defining the minimally competent candidate, they are being provided with data about a group that someone else has defined as borderline. Plake, Melican & Mills (1991) recommend providing item difficulties for the whole group.

Table 4
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	50	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

Adapted from Englehard and Cramer (in press), presented in Cizek (1996b).

The presence or absence of an answer key may have a differential effect of ratings of familiar and unfamiliar content areas (Plake, Melican and Mills, 1991). Although training can help to equalize knowledge, some judges may still answer items incorrectly. In this case, discussion between judges who answered correctly and judges who answered incorrectly can be helpful; this can help to refocus judges from their specialized expertise to the broader domain of knowledge (Chang, Dziuban, Hynes & Olson, 1996).

"[The Angoff method] 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 assessment."

Advantages of the Angoff method.

1) It is not difficult to explain and data collection and analysis are relatively simple (Mills & Melican, 1988).

2) Its properties have been extensively researched.

Jaeger's method

"Jaeger (1982) developed another item-based procedure, similar to that initially suggested by Angoff (1971)... Jaeger's procedure requires sampling from each population with an informed, legitimate interest in the outcome. 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). 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)...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. Because the Jaeger method has been introduced more recently, it has received comparatively less scrutiny than the Angoff, Ebel or Nedelsky approaches."

Advantages of the Jaeger method.

1) "Its explicit recognition of the fact that various constituencies have a stake in the results of the standard-setting process."

2) "It does not require judges to conceptualize the minimally competent examinee. Thus, groups that have an interest in the test results, but do not have

sufficient contact with the examinee population to make judgments about the test performance of minimally competent examinees may be included as judges" (Mills & Melican, 1988).

Disadvantages of the Jaeger method.

1) "The wording of the question asked of judges implies that passing status could be denied on the basis of a single item. Even though the calculation of a cutoff score using this method does not place such weight on individual items, judges may be reluctant to respond affirmatively to the rating task given its wording (Mills & Melican, 1988).

2) As in the Nedelsky procedure, participants can only make probability choices of 0 or 1 (Berk, 1986)

3) According to some research, it produces somewhat less reliable standards than other item-based approaches (Cross, Impara, Frary & Jaeger, 1984).

Challenge to test-centered methods

Recently, a challenge has been made to one of the assumptions underlying all the test-centered methods, namely, the assumption that judges are capable of accurately performing the judgment task. While specifically motivated by the National Assessment of Educational Progress (NAEP) and to the Angoff method used by NAEP, the criticism challenges to all test-centered methods.

"A report of the National Academy of Education studied implementation of a modified Angoff approach used to set standards for 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 candidates, may be impossible for participants to acquire or 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).

Cizek notes that because hypotheses such as those of Shepard et al. have not received much empirical attention to date, it is likely that item judgment methods will continue to be popular in the near future. A very new approach described in the next section, the "Bookmark Approach" (Mitzel, Lewis & Green, 1996) may constitute a response to these hypotheses. Berk's General Eclectic Method, described later in this chapter, also avoids conceptualizations of minimally competent candidates.

"The Bookmark Approach"

Mitzel, Lewis and Green (1996) believe that criticisms such as those of Shepard et al. (1993) are justified based on cognitive theories of judgment and decision-making. The authors introduced and implemented an IRT-based standard-setting method, which was used on a new assessment of reading, language, math, science and social studies developed by CTB/McGraw Hill, called TerraNova.

In the Bookmark Approach, judges set cut points through use of special test booklets in which the items are presented in order of IRT ability scale location. The location of a multiple-choice item is defined as the point on the ability scale at which a student would have a .67 probability of success (with guessing factored out). This level was chosen because it is a point at which most but not all students would be answering successfully. For constructed response items, each item has more than one location in that each possible score point (i.e., getting part of the item correct) for the item has a unique location. The location of a score point is defined as the point on the ability scale for which students have a .67 probability of achieving at least that score.

Prior to rating items, judges receive training regarding the fact that they might perceive some disordinality among the items, i.e. that some later items might seem easier to them than earlier items. This would occur because of variations in the criteria the judge considers in order to judge item difficulty. Following training, the judges' task in setting cut scores is to place bookmarks in the ordered item books at the positions such that "on the whole" the items preceding the bookmarks are items reflecting content that the given performance level student should know and be able to do with at least 2/3 likelihood.

Before setting cutpoints, judges develop initial descriptions for each performance category of what students classified at a given level typically know and are able to do (exact details of how the discussion proceeds are provided in Mitzel, Lewis & Green 1996). The purpose of concentrating on content rather than item difficulty was to ease the later task of Because an additional goal of the standard-setting process was developing descriptions for each performance category of what students classified at a given level typically know and are able to do, the judgment task was designed to focus judges' attention on item content rather than on their judgments of item difficulty (as occurs in the traditional modified Angoff procedure). Because the test booklets are used to create initial descriptions, approximate cut point locations are implied by the initial definitions.

The standard-setting process consists of three rounds. In the following description, three performance levels are used as the example, Below Proficient,

Proficient, and Advanced, although in the actual TerraNova test, five levels were established. To establish three performance levels, two cut scores are needed.

In the first round, judges use the group's initial descriptions to place two bookmarks. It was found that in placing the bookmarks, it was easy for judges to first identify an interval of items in which the bookmark clearly belongs, then consider the items more carefully to identify the exact placement of the bookmark. Because of "local disordinality", judge are asked to consider items well beyond the first item they perceive as being "too difficult" for the given performance level student.

In Round 2, the cut scores for the Proficient (highest) level assigned by each judge are displayed. The items between the scores assigned by different judges are discussed (exact details of how the discussion proceeds are provided in Mitzel, Lewis & Green 1996, for example, the judge whose judgment is an "outlier" first defends his or her classifications). It is hoped that discussion will elicit a shared perspective. Then the Advanced cut-lines are considered in a similar manner.

In Round 3, the judges are presented with: (a) the placement of each of the judges' bookmarks for the two cut scores; (b) the average placement for each cut score, and (c) impact data, i.e. the percentage of students falling in each category based on the group bookmark (optional). The judges discuss the final cut in terms of the implications toward what the given performance level student should be likely to know and be able to do. Following discussion, they may adjust their bookmark locations a final time.

After the final cut scores are determined in Round 3, the initial descriptions are modified to their final form based on the tendencies exhibited by items at each performance level. For example, the items between the proficient and advanced cut lines represent items which some of the proficient students are likely to know and be able to do but other proficient students do not.

Advantages of the Bookmark Approach (according to Mitzel, Lewis and Green, 1996).

1) Cut scores are based on a comprehensive understanding of the test content;

2) Defensible performance level descriptions that are a natural outcome of setting the cutpoints.

3) Committee members using the bookmark process found it to be "rational, interesting, and professionally enriching. The discussions engaged in by the participants "brought out the processes and knowledge required of the students to successfully respond to the items; one group discussed each item at length one by one in increasing order of difficulty. When they reached the last item, they discussed the gray zones and placed the group bookmarks within 30 minutes...Nearly 100% of the

participants indicated they would like to be part of future standard settings of this type."

4) Essentially no data entry.

5) Time efficiency.

Disadvantages of the Bookmark Approach.

Two potential problems with the Bookmark Approach, but not unique to it, are noted by the authors:

1) If the test is not composed of items that span the full range of ability for which standards are being set, a floor or ceiling effect may be encountered.

2) When writing descriptions, the items which appear between the proficient cut score and the advanced cut score were erroneously perceived by some groups to be "the proficient items", that is, the items that the proficient students should be likely to know and be able to do. Actually, however, the items which all proficient students are likely to know and be able to do are at locations *below* the proficient cut score (i.e. the items preceding the proficient bookmark). This problem can be avoided by instructing the judges about the matter.

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." Two examinee-centered methods will be discussed briefly in this section. However, the properties of examinee-centered methods have not ben studied very much (Kane, 1994), and they have been used less commonly, so less information is available about the examinee-centered methods.

"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. Actual

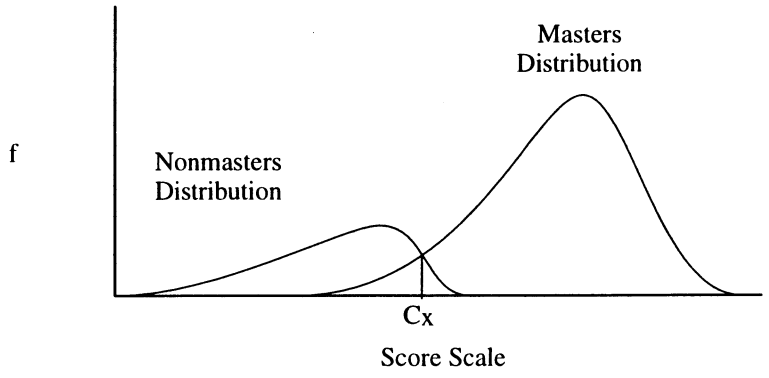
performances of real people are judged, as opposed to eliciting estimates about the probable performance of a hypothetical group."

Contrasting Groups Method

"The contrasting groups method was described under another name by Berk (1976)...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).

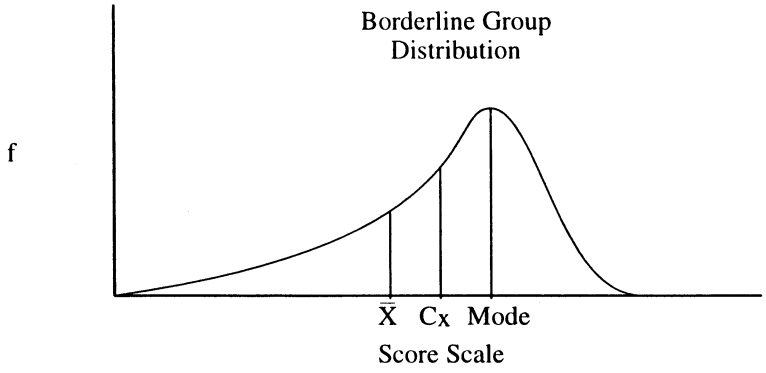
"A variation of Berk's method involves asking participants, who have knowledge of both the examinee population and the required knowledge or skill level, to classify examinees as either competent or not competent. Livingston and Zieky (1982) recommend plotting the percentage of test takers at each score level who are judged to be competent." One procedure they recommend for deriving a passing score is to "select a point that minimizes the overall impact of errors of classification. Figure 1 illustrates a passing score obtained using the contrasting groups method, with the cutting score indicated as C_x."

Figure 1. Contrasting Groups Method



(Cizek, 1996b)

Figure 2. Borderline Group Method



(Cizek, 1996b)

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. Figure 2 illustrates a passing score obtained using the borderline group method."

Advantages of examinee-centered methods.

1) Because the cutoff scores are directly related to actual test performance, there is usually little need to adjust the cutoff score (Mills & Melican, 1988).

2) "The methods are easy to explain, and data collection and analysis are straightforward" (Mills & Melican, 1988). However, Kane (1994) states that the methods are often difficult to implement.

Disadvantage of examinee-centered methods.

"The pool from which judges may be selected is limited. Because these methods require both knowledge of the test content and of the achievement of the examinees, judges most often are limited to course instructors" (Mills & Melican, 1988).

Disadvantages of the contrasting groups method.

1) "A cutoff score established by the contrasting groups method may simply preserve the status quo (Mills & Melican, 1988).

2) "Because the establishment of the cutoff score relies in large part on the performance of individuals in the tails of the two distributions (high-scoring nonmasters and low-scoring masters), the effect on the cutoff score of misclassification of individuals in the initial rating task can be substantial. A relatively small number of

high scores in the nonmaster group can raise the cutoff score, and low scores in the master group can lower it" (Mills & Melican, 1988).

Disadvantages of the borderline group method.

1) "It is often difficult to identify a borderline group of sufficient size.

2) "Jaeger (1989) 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'.

3) "Jaeger (1989) also 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." (Reasons for making adjustments to cutoff scores are discussed further in Chapter 3).

Three of the most commonly used compromise models, the Beuk, deGruijter, and Hofstee models, are described in the following sections. Little comparative work has been done on these models, however (Cizek, 1996b).

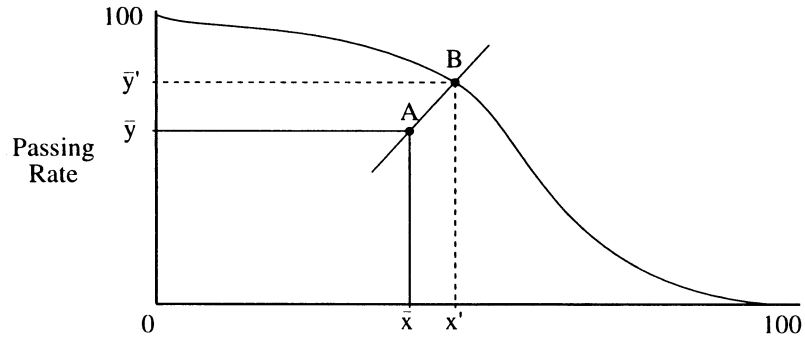
Beuk's Method

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

"Figure 3 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 percentage 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 ($\overline{X}$ ') by the number of items in the examination.

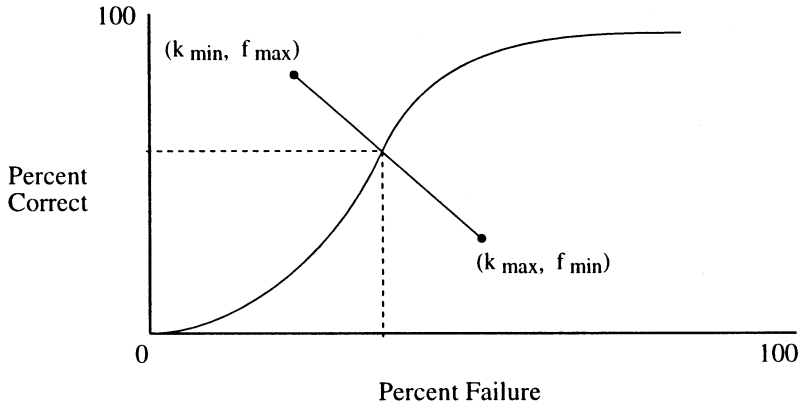
Figure 3. Beuk Method



Where: $\bar{x}$ = mean expected percent correct (over judges)
$\bar{y}$ = mean expected passing rate (over judges)
$\bar{x}'$ = adjusted percent correct
$\bar{y}'$ = adjusted passing rate
S_X/S_Y = slope of line AB

(Cizek, 1996b)

Figure 4. Hofstee Method



(Cizek, 1996b)

Advantages of the Beuk method.

1) "The computations are simple and the application appears logical; the more the judges agree on their estimates on one dimension (i.e. relative or absolute) the smaller the degree of adjustment on that dimension will be" (Mills & Melican, 1988).

2) "In addition to the information typically collected in the first phase of a study, it requires only that judges provide an estimate of the passing rate" (Mills & Melican, 1988).

Disadvantage of the Beuk method.

"A potential drawback of this method is that judges may have difficulty specifying expected passing rates if they have not had experience with the entire range of performance in the examinee group. For example, teachers from highly achieving districts may have substantially different views of student competence than those in less highly performing districts." (Mills & Melican, 1988).

Hofstee's method

"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 scores. The lower portion of Figure 4 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). 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 (Cizek, 1996b)." The first connects two points, that represented by the lowest possible cutting score and the highest possible failing rate and that represented by the highest possible cutting score and the lowest possible failing rate. The second is based on the empirically determined failure rate for each possible test score. The abscissa value for the point of intersection of these two lines serves as the compromise cutoff" (Geisinger, 1991).

Advantage of the Hofstee method.

It is simple (Mills & Melican, 1988).

Disadvantages of the Hofstee method.

1) The line segment for any judge may not intersect the test score distribution (Mills & Melican, 1987; Jaeger, 1989). In this case, the method does not yield a cutoff score unless the line segment is extended beyond the limits set by the judge. This action is inappropriate because the Hofstee method has the implicit assumption that the only acceptable compromises lie between the extreme values established by the judges" (Mills & Melican, 1988).

2) "The apparent ease of collecting the absolute information (i.e. not having to perform one of the absolute standard-setting studies) can lead to a cursory view of test content and to estimates that do not fully incorporate the difficulty of the test (Mills & Melican, 1988).

deGruijter's Method[5]

"DeGruijter (1985) noted that, although the Beuk method adjusts for variability within the judge group, it does not account for any uncertainty that a particular judge may have. DeGruijter's method allows the incorporation of uncertainty estimates on the relative and absolute scales into the compromise adjustment.

"In addition to the data collected in the Beuk method, the deGruijter method requires that each judge provide estimates of his or her uncertainty with respect to the

---

[5] The information in this section is quoted from Mills & Melican, 1988.

true value of the failing rate and the true value of the cutoff score. The amount of adjustment made to the estimates is a function of the ratio of these uncertainty estimates.

"Like the Beuk method, the deGruijter method involves plotting the cumulative frequency distribution. The point associated with the estimated absolute cutoff score and the estimated passing percentage is plotted for each judge. Using the uncertainty estimates, a family of ellipses is generated about this point. One ellipse just touches the cumulative frequency curve, and the cutoff score associated with this ellipse is the 'compromise' cutoff score.

Advantage of the deGruijter method.

"The only compromise method that incorporates judges' confidence in their ratings into the adjustment of the cutoff score."

Disadvantages of the deGruijter method.

1) "Computationally complex and difficult to explain."

2) "Quantification of the uncertainty estimates is difficult and appropriate techniques for collecting uncertainty data have not been developed" (Mills & Melican, 1987).

Performance Assessment

A special issue of *Applied Measurement in Education* (1995) was devoted to the issue of standard setting for performance assessments (e.g., portfolios). The three methods discussed there, judgmental policy capturing (JPC), extended Angoff procedure, and dominant profile method (DPM), will be introduced very briefly.

JPC focuses on the process or strategy used by judges to arrive at their ratings (Hobson & Gibson, 1983). "The goal of JPC (Jaeger, 1995) is to describe the policy that represents the panel's judgments via a statistical equation that defines the captured policy....JPC consists of two stages: (a) judges independently rate multidimensional profiles of performance to specify an exercise score of *poor* (1) to *excellent* (5) and (b) judges independently rate hypothetical candidates with specified profiles of exercise scores from *novice* (1) to *highly accomplished* (4). The latter ratings weight the importance of the exercise scores" (Berk, 1995).

Hambleton and Plake's (1995) extension of the Angoff procedure involves revision of two judgment tasks as accommodations to performance-based assessment: "(a) specification of expected scores for just barely certifiable candidates on polytomously scored exercises and (b) weights to reflect the relative importance of scoring dimensions at the exercise level and the exercises themselves."

In applying the dominant profile method (Putnam, Pence & Jaeger, 1995), "judges explicitly create and state decision policies across the entire assessment package rather than render decisions exercise-by-exercise. First, judges state decision policies and construct bottom-line profiles of just barely certifiable candidates. Then, one week later, judges are given feedback on their recommendations and revise their original ones. Last, judges are requested to respond to a set of challenge profiles that challenge their proposed policies. The purpose of the DPM is to assess whether the judges can agree on a single bottom-line policy (the dominant profile) that matches their recommendations to certify or not certify a candidate with a particular profile of scores" (Berk, 1995).

None of these new methods was completely successful in describing a distinctive policy of judges to arrive at a performance standard, but they do represent valuable research toward grappling with the complexities that performance assessment brings to standard setting (Berk, 1996).

General Eclectic Method (GEM)

"Two major testing practices that emerged in the decade of the 1990s – the use of polytomous item formats and multiple cut scores–have stimulated a new generation of standard-setting methods" (Berk, 1996). Berk combined "the best strategies from the decade of the 1980s with a proven track record...with the most promising new techniques into a General Eclectic Method (GEM) for standard setting." Berk outlines the GEM as a ten-step process:

"1. A broad-based sample of judges defines achievement levels based on consensus.

2. A subsample of the judges (content experts) provides amplified, explicit behavioral descriptions of achievement levels based on consensus.

3. Judges select anchor or prototypic unscored items (test-centered) or scored items or work samples (i.e., portfolios) (examinee-centered) at upper and lower ends of achievement-level categories based on consensus.

4. Judges independently match unscored items (test-centered) or scored items or work samples (examinee-centered) to achievement-level categories based on behavioral descriptions and anchors.

5. Judges independently rate the importance, difficulty or complexity of each objective, outcome, or dimension (optional).

6. Judges are given feedback on their individual and the panel's decisions (from Step 4) plus meaningful performance data and requested to independently revise their initial decisions.
7. Judges discuss their revised item or work sample classifications and importance ratings (optional) without pressure to reach consensus.
8. Judges render their final independent revisions of their classifications and ratings (optional) based on the discussion (Step 7) as well as accumulated insights.
9. Determine the judges' decision policy rule from the designated weights finally assigned to the objectives, outcomes, or dimensions (Step 8; required only if Step 5 was executed).
10. Compute the cut-scores for the total test or by objective or dimension based on the decision policy rule (Optional Step 9) and the mean or median item or test scores from the judges final classifications at each achievement level (Step 8) in terms of the chosen test score scale.

The GEM has not been put into use and critiques of it are not yet available.

Comparison of methods

Little comparative work has been done on the compromise models (Cizek, 1996b) or the examinee-centered models (Kane, 1994). "Among [the methods discussed above], the Angoff method, or one of its many variants, is probably the most widely used for so-called "high stakes" educational tests and licensure and certification contexts" (Cizek, 1996b). "Research on the Angoff method has suggested that it provided easy-to-obtain and acceptable results in many situations.... Berk (1986) concluded that 'the Angoff method appears to offer the best balance between technical adequacy and practicability'."

"Although the evidence is somewhat mixed, the Angoff method seems to have relatively small standard errors for the passing scores and, because reliability is a necessary condition for validity, the sizes of the standard errors are relevant to questions of validity. Note, however, that the reliability issue is not decisive, for two reasons. First, although reliability is necessary for validity, it is not sufficient. The factors that yield relatively small standard errors for the Angoff method, which could include a tendency for judges to restrict the range of their ratings or to pay too much attention to item statistics, might not be associated with improved validity for the

intended interpretation. Second, it is always possible to decrease the standard error and therefore improve reliability for any method by increasing the number of judges, and/or items ... used in standard setting" (Kane, 1994).

Recent criticism of test-centered methods resulted in the Bookmark Approach, which, though intriguing in its provision of rank-ordered difficulties, thus far has not been used except by its authors.

CHAPTER 2: JUDGMENTS

All standard-setting methods involve judgments (Jaeger, 1989); therefore, any issues connected to the judgment process are crucial to the validity and reliability of the standard-setting process. The issues discussed in this chapter include: how to select qualified judges, how many judges should be selected, how to train the judges, and how to evaluate and use their judgments in determining the passing score.

Selection of judges

Expertise

Just one standard of the *Standards for Educational and Psychological Testing,* Standard 6.9, addresses the issue of judges' qualifications, and it addresses only the issue of documentation, not the selection itself: "When cut scores are based primarily on professional judgment, the qualifications of the judges should be documented" (p. 43). The guidelines provided in the *Standards* are not useful enough for selecting the judges, so other sources of information must be used (Jaeger, 1991).

The literature on the nature of expertise, while not written for the purpose of judge selection for standard setting, is still relevant. Chi, Glaser and Farr (1988) characterize experts in the following ways (listed in Jaeger, 1991):

1. Experts excel mainly in their own domains, rather than because of global qualities of their thinking.

2. Experts are able to perceive large meaningful patterns in their domain of expertise because of their organized knowledge base, rather than because of global qualities of their thinking.

3. Experts are able to perform rapidly and accurately in their domain of expertise. This speed is based both on their knowledge base and on the capability of arriving at a solution without extensive search of their knowledge base.

4. Experts see and represent a problem at a deep level, while novices tend to represent problems at a superficial level.

5. Experts spend time analyzing a problem qualitatively. They build a mental representation of a problem from which they infer relationships that define the problem situation and then they add constraints. Adding constraints reduces the "search space" for solutions.

6. Experts have strong self-monitoring skills. They seem to be more aware than novices of when they make errors, why they fail to comprehend and when they need to check their solutions.

7. Experts are more accurate than novices at judging problem difficulty (Chi, Glaser & Rees, 1982).

These descriptions suggest that expert judges will be found among people who have direct experience in the particular domain that is being tested. However, it is not the case that everyone who has direct experience will make a good judge.

Broad representation of interested groups

"All important points of view should be represented on the panel of judges" (Livingston & Zieky, 1982, p. 16). Audiences "qualified" to have a say in the process, for example, for a high school diploma test, would include administrators, parent representatives, and representatives of business and government. In other cases, very few people will be qualified, for example, for a test of certification in a particular sub-specialty of medicine (Cizek, 1996a). The specific qualifications needed in the judges will depend on the purpose of setting the standard in the first place, the type of decision to be made using the passing score, and political considerations (Jaeger, 1991; Cizek, 1996b).

Sometimes a conflict of interests can arise between a desire to have a broad representation of interest groups among the judges and the requirement that judges be qualified. Cizek (1996b) suggests empaneling stakeholders who additionally possess relevant knowledge of the area and population assessed. If a choice must be made, the decision between broad representation and homogeneity of participants will have to be made based on the priorities of the particular standard-setting application (Cizek, 1996a).

Number of judges

Two approaches have been suggested for determining how many judges should be used for a standard-setting panel: (a) the size of the sample of judges should be such that the standard error of the mean (or median) recommended standard is small compared to the standard error of measurement of the test (Jaeger, 1991); and (b) use of as many judges as possible (e.g. Cizek, 1996b). These approaches will be described in turn.

Jaeger (1991) cautions that using as many judges as possible can be wasteful or insufficient. One can calculate the standard error of the mean of the judges' estimates, (assuming the judges are sampled randomly, independently and with equal probability):

$$\text{St. error of mean} = (\text{SD}) / \sqrt{n} \qquad (1)$$

where n is the number of judges in the sample, and SD is the standard deviation of the test standards recommended by individual judges.

Jaeger first considers, but apparently rejects, a criterion that the standard error of the mean recommended test score not exceed 0.5 test items. This criterion results from a specification that no examinee (or very few examinees) with test scores within one point of the test standard that would have been recommended by the total population of judges, would be misclassified due to estimation errors that arise from sampling of judges. The value 0.5 reduces the estimation error to no more than one test item for 95 percent of the samples of judges that might be drawn, assuming a fixed sample size. This might seem to be a reasonable specification, however, a theoretical example using this criterion led to a large, probably impractical number of judges being necessary (e.g., 87 judges).

An alternative criterion that Jaeger suggests for determining an acceptable size for the standard error of the mean recommended test standard is comparing it to another major source of error, the standard error of measurement of the test. If errors due to the unreliability of the test (the standard error of measurement) and errors due to sampling of the judges are considered to be independent, a standard error that reflects both sources of error is given by the following:

$$\text{St. error} = \sqrt{(\text{St. error of Mean}^2 + \text{St. error of Msmt}^2)} \qquad (2)$$

This equation can be used to show that "the overall standard error is only three percent larger than the standard error of measurement, provided the standard error of the mean is no larger than one-fourth the size of the standard error of measurement of the test.... One might thus argue that a sufficiently large sample of judges is one that will result in a standard error of the mean recommended test score that was no larger than one-fourth the standard error of measurement of the test for which a standard is sought ." In a practical example, the standard error of measurement on the National Teacher Examination's General Knowledge Test (Educational Testing Service, 1984), was reported as 5.2 items. If the standard error of the mean recommended test standard was one-fourth of this value or 1.2 items, the overall standard error would increase only 3.08% to 5.36 items. Using Equation (1) (rearranged), and a value of 4.65 for the standard deviation (this was the median value found in nine studies of the modified-Angoff standard-setting method), the number of judges required would be

$n = (4.65/1.3)^2 = 12.79$ or 13 judges, rounding up.

Cizek (1996b) criticized Jaeger's method for applying only (as Jaeger himself notes) when the population of judges is sampled randomly, independently, and with equal probability, conditions which are frequently violated. He therefore advises to utilize as many judges as practical, given available resources.

Actual studies of varying the numbers of judges have yielded different results. Livingston and Zieky (1982) reported a successful effort with as few as five participants, as did Norcini, Shea and Grosso (1991). However, Norcini, Shea and Grosso note that their participants also participated in the test development process, which may have improved their consistency in the standard-setting. Smith, Smith, Richards and Barnhardt (1988) found that variability was still large with 10 participants. However, consensus building can be difficult as the number of judges exceeds 20 (Berk, 1996).

Training

If the panelists in standard-setting studies are, as they should be, chosen to represent all appropriate groups in the profession relevant to establishing the cutoff scores for the test, they will bring a diversity of knowledge, training, and opinions about the test to the rating session. Also, the panelists are generally unfamiliar with the task they are expected to perform. The logical solution to these problems is training (Mills, Melican & Ahluwalia, 1991). In contrast to this model stance, however, most actual standard-setting studies do not report orientation activities for judges, and, as best as can be determined, most studies include little structured training (Reid, 1991).

Although there is little direct evidence that training has an effect on judges' behavior, Reid (1991) believes it is best to adopt a conservative approach and assume that lack of training will have a detrimental effect on the quality of the standard-setting data. Reid suggests individualized, possibly computerized training for judges with different backgrounds.

Mills, Melican and Ahluwalia (1991) provide practical guidance for three steps in the training process that should take place prior to training judges in how to rate items: explaining the process, setting the context of the meeting, and (if applicable) developing a common definition of minimal competence. The goal is to create a situation in which differences in ratings result solely from differences in perceptions of item difficulty only, and not from differences in the definition of minimal competence, the usefulness of the test, or other factors. They suggest that arriving at a definition of minimal competence should take half a day using their method, but that this depends on the domain and the prior experience of the judges.

Explaining the process

In order to avoid confusion and allay concerns of participants, the following issues should be discussed at an initial meeting:

1) An overview of the standard-setting process;

2) A discussion of the judges' role and the reasons they were chosen;

3) An explanation of the data collection form;

4) A chance to voice questions and concerns about their adequacy to participate.

Setting the context of the meeting

This should include:

1) An explanation of the decision to be made on the basis of the test results; this is important because item ratings could depend on what the decision certifies, e.g., merely theoretical knowledge of subject matter or practical, on-the-job ability.

2) A discussion of the steps taken to develop the content specifications and test items. Judges, may, as experts in their fields, have concerns about the test as it stands, but their task is limited to standard-setting, and does not include rewriting the test.

3) A description of where judges' input will be used (e.g., only as a recommendation or in determining the final cutoff score).

Defining minimal competence

A specific definition for the test must be developed, because a generic definition such as "the minimal level of knowledge and skills required" does not take into account, for example, the variety of skills being tested and the possible compensatory effect of strength in one area on deficiencies in another. Also, participants may confuse entry-level minimal competence with minimal competence of practitioners.

Mills, Melican and Ahluwalia (1991) provide a detailed example of the process of arriving at a definition of minimal competence for a geology test. The process takes place in a group setting. A summary of the process they describe is as follows: Begin with a global definition of the skills routinely required for practice (knowledges, skills, and abilities of the profession). The global definition can be based on the test specifications. Then, develop refinements that address minimally acceptable levels for the beginning practitioner. For example, judges can answer the question, "What types of mistakes are forgivable errors for a minimally competent candidate?" Once this is done, judges should develop a concise definition of minimal competence.

Fehrmann, Woehr and Arthur (1991) found that "frame of reference training", meaning either providing a definition of minimal competence in the specific domain of

the test, or having the group arrive at a definition, led to greater interrater consistency than proceeding with the Angoff method without such a definition.

Training judges to rate items

The last step in training involves training judges to apply the definition of minimal competence. Reid (1991) recommends providing practice in the steps of the particular standard-setting method to be used. Practicing should be done on training items rather than on the first items in the actual set, for two reasons. First, it is less of an ego blow to judges to change a rating that is out of line on a training item than to change a rating on an actual item. Second, working on training items blocks the tendency of some judges to forge ahead at all costs once they are working on actual test items.

Ambiguities of the methodology of the standard-setting method should be pointed out to the judges, for example, whether or not to factor in relevance in the Angoff method. Also, it can help judges to know about factors affecting item difficulty and difficulty of rating items. For example, judges may think erroneously that item format affects item difficulty more than it actually does. Another example is that it has been found that negatively phrased items are hard for judges to rate (Melican & Thomas, 1984).

CHAPTER 3: VALIDITY

"Validity is a property of the interpretation assigned to test scores and not a property of the test itself or of the test scores (Messick, 1988, 1989)....The appropriateness of the proposed interpretation of the passing score "depends on two assumptions: (a) that the passing score corresponds to the specified performance standard, in the sense that examinees with scores above the passing score are likely to meet the standard and examinees with scores below the passing score are not likely to meet the standard; and (b) that the specified performance standard is reasonable given the purpose of the decision" (Kane, 1994). Kane terms these assumptions, respectively, the *descriptive assumption* and the *policy assumption.* The distinction between the descriptive and policy assumptions is useful for categorizing the three sources of validity evidence discussed in this section: procedural, internal, and external. Procedural evidence and internal criteria are most useful for supporting the descriptive assumption, and external criteria are most useful for providing evidence for the policy assumption (there is overlap, however). According to Kane, the choice of passing score, ultimately, is fairly arbitrary; the best we can do is to show that the passing score and its associated performance standard are not unreasonable.

Procedural Evidence

Although implementation of the best available procedures does not guarantee the appropriateness of the passing score, poor procedures do cast doubt on the appropriateness of the passing score (Kane, 1994).

In part of the following discussion, Cizek's (1996a) guidelines (see Introduction, Table 1) will be followed; many of his guidelines are efforts to promote procedural validity.

Defining the purpose for setting standards

Before choosing a standard setting method, it is vital to consider whether it is necessary or useful to employ a passing score altogether in a particular case (e.g., Glass, 1978; Kane, 1994), as opposed to some other method of testing such as norm-referencing. "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 standard-setting process" (Kane, 1994). Although this step may be straightforward, it is often overlooked (Cizek, 1996a). In one example of a rationale for standard setting, the American Board

of Medical Specialties stated the intent to "provide assurance to the public that the diplomate has successfully completed...an evaluation to assess knowledge, experience, and skills requisite to the provision of high quality medical care" (Langsley, 1987). More detailed rationales could be provided, e.g., in this example, what constitutes high quality knowledge, skill, care and so forth (Cizek, 1996a).

Defining relevant constructs

"Clear definitions are a critical consideration in conducting judgmental and empirical passing score studies. For example.... the ability of participants to conceptualize the hypothetical person–a construct–referred to in the Angoff procedure is critical to the success of that method.... In practice, participants will usually need to make repeated reference to a formal, written summarization of the attributes and performance indicators that define the construct of interest in order to maintain conceptual fidelity and to implement the procedure faithfully."

Connecting the purpose and method of setting standards

Potentially conflicting purposes can exist in a standard setting process. For example, a school system might be interested not only in awarding diplomas to qualified applicants, but also in awarding diplomas to as many applicants as possible so that young adults are not handicapped when job-hunting. Whatever the goals of the testing program are, they should be clearly stated at the outset so that they can be considered when the standard-setting method is selected and so that the passing score will have meaning (Cizek, 1996a).

Connecting characteristics assessed and method

"The method selected for deriving a passing score should vary according to the knowledge, skills, and abilities assessed, and the contexts in which they are assessed. For example, a procedure such as Nedelsky's is appropriate for multiple-choice items (where the characteristic assessed is knowledge), though not for polytomously scored essay items (where the characteristic assessed is the ability to produce a writing sample) (Cizek, 1996a).

Describing the standard-setting method selected

"In describing the conduct of a passing score study, it is essential that the method chosen for deriving the passing score be clearly described and referenced. For the methods [discussed] above [except the Bookmark Approach], ample description can be found in the professional literature. The literature also contains specific

references that address the technical adequacy of the method chosen; when available, information supporting the professional acceptability and technical adequacy should be provided.

"More extensive methodological description is warranted if the standard-setting procedure is a variation of a documented procedure or if the method is obscure, innovative, or poorly documented in the literature. In these cases, a detailed description of the procedure, with a sound and compelling rationale for its use, should be provided" (Cizek, 1996a).

Description of the participant group

"Adequate documentation of a standard-setting process should present a description of the participant group such that there is a discernable match between the characteristics of the group and the judgments that must be made....Both the number of participants and their ability to make the required judgments are relevant information (Cizek, 1996a).

Evidence of task comprehension

"Even the most carefully designed and implemented training procedures do not guarantee that all participants fully comprehend the task. Information can, and should, be gathered to verify that participants assimilated the training; such information can serve as an additional source of validity evidence. For example, monitoring of group activities and recording of group discussions can yield anecdotal evidence that participants understand the task. It is also a good idea to have a written record of the characteristics of the hypothetical "borderline proficient" or "minimally competent" student about whom judgments will be made; the list of characteristics can serve as a reference throughout the standard-setting activities, as well as provide documentation of the characteristics for the standard-setting report" (Cizek, 1996a).

Appropriate use of information

Evidence that participants actually used the information (e.g., normative data)–and used it appropriately–is...desirable (Cizek, 1996a).

Feedback from judges

"Post-standard-setting questionnaires should also be considered as a means of surveying the extent to which participants report that they have understood the training, felt that their opinion was considered, were confident in applying the methodology, and believe the resulting standards to be reasonable and defensible. If implemented, both

the qualitative and quantitative survey results should be included in the standard-setting report" (Cizek, 1996a).

Documentation

"In documenting a standard-setting study, it is advisable that all aspects of the procedure used in actually implementing the chosen procedure be explicated in detail. The following list represents the minimum in terms of the kinds of information that should be included: the number and manner of selecting participants; the qualifications of participants, the qualifications of those designing and implementing the methodology; the materials used; the script or actual verbal instructions given to participants; key frameworks or conceptualizations developed by participants (e.g., lists of expected student proficiencies or descriptions of minimally competent performance); the timeline, schedule of events, and actual agenda followed....

"Departures from intended procedures do not necessarily diminish the defensibility of the resulting standard. In many cases, failing to implement the procedure as planned may serve to support the validity of the resulting passing score. For example, it is not unusual for a standard-setting study to include additional or refresher training to participants for cases in which there is concern about participants' understanding of the task, consensus on key conceptualizations, or agreement on policy implications. In any case, deviations from intended procedures should be noted and carefully explicated in the documentation of a standard-setting study" (Cizek, 1996a).

Internal Criteria

"The emphasis in the design of internal validity checks is on the consistency of different sets of results derived from the study. Consistency in the results does not provide compelling evidence for the validity of the proposed interpretation of the passing score, but it does provide support for validity (Kane, 1994).

Factors affecting consistency of judgments

Content knowledge. Various studies of the effects of content specialties of judges on their decisions have provided discrepant results (Cizek, 1996). Also, all of these studies compared content knowledge within a limited domain, e.g. judgments of social studies specialists versus those of mathematics specialists, all of whom were professors in teacher education programs (Plake, Impara & Potenza, 1994); no studies compared content specialists with stakeholders. In one actual standard-setting effort,

the results of varying the group of judges were also equivocal. In the 1990 NAEP mathematics assessment, noneducators and educators did not come to a consensus on standards; it was suspected that this was because the different groups used different criteria. However, repeating the standard setting process with panels consisting mostly of teacher educators did not lead to more homogeneous ratings (Bourque & Hambleton, 1993).

Chang, Dziuban, Hynes & Olson (1996) found that judges tended to set lower, and less consistent standards for items that they themselves could not answer; they suggest training before items are rated in order to ensure that all judges are competent in all parts of the domain.

Group psychology. One disadvantage of group discussions of ratings is that social psychological effects can result, for example, a dominant personality taking over the discussion (Fitzpatrick, 1989). However, the benefits of such discussions are considered to outweigh this disadvantage (e.g., Kane, 1994). Several suggestions have been made for mitigating potential social effects. Fitzpatrick suggests asking group members to discuss their points of view about items without revealing what specific rating they assigned to an item. Another possibility is for judges to revise their ratings after seeing the ratings provided by others, but without discussion (Plake, Melican & Mills, 1991).

Idiosyncratic perceptions of skills required of the minimally competent candidate on the part of a particular judge (Mills, Melican & Ahluwalia, 1991).

A judge did not understand the judgment task. A judge may even express serious doubts about his or her own data. Although training ideally addresses these matters, they can still remain problematic (Plake, Melican & Mills, 1991).

Personal stake in the outcome of the standard setting, e.g. a family member of a judge took the test (Geisinger, 1991).

Disapproval of the test by a rater. Although training addresses this issue as well (Mills, Melican & Ahluwalia, 1991), it can nevertheless affect the seriousness with which the rater approaches the task (Plake, Melican & Mills, 1991). For example, ratings made by individual judges may appear inappropriate, e.g. ditto marks indicate the same percentage for every question, or the item evaluations lead to a passing score below chance level (Geisinger, 1991).

One approach to solving inconsistency problems is periodic retraining during the rating task. This can insure that judges do not drift from the definition of the minimally competent candidate, and that they scrutinize each item. However, retraining can cause rater fatigue due to the increased time required or the disruption of

concentration. Also, social influence is again a possibility (Plake, Melican & Mills, 1991).

Measuring consistency

Intrajudge inconsistency. "Intrajudge inconsistency arises when the judge specifies probabilities of success on the items which are incompatible with each other, and consequently, imply different standards. An example is an Angoff judge assigning a low probability of success for a borderline student on an easy item but a large probability on a difficult item" (van der Linden, 1982). "The [van der Linden] consistency index of judges' recommendations is computed by examining deviations [by the judge] from estimates that would result if the IRT model were applied to the item performances of an examinee with the specified ability" (Cizek, 1996a).

Consistency of the group judgment with actual item difficulty. The van der Linden index can also be used to detect whether the empirical estimates of the proportion of marginal examinees answering an item correctly are consistent with the judgments made by the judges *as a group* of the probability that a minimally competent examinee would answer the item correctly. If inconsistency occurs on too many items, the passing score may be set too high or low. According to Kane, "in either case, we have some evidence that the item characteristics that the judges are using to evaluate item minimum passing levels are different from the item characteristics that are determining the difficulty of the items for examinees, and this would cast doubt on the interpretability of resulting passing scores in terms of the performance standard".

Ratings should reflect realistic expectations of candidates. On the other hand, a judge [or judges] may have rightly "upheld the standard" in rating an item as one which minimally competent students should be able to answer. That is, standards should not be discounted due to actual poor performance of examinees. Thus, whether to implement the criterion of being realistic will depend on how discrepancies from this criterion are viewed (Reid, 1991).

Interjudge inconsistency. This may be measured using Jaeger's (1988) modified caution index (MCI). The assumption underlying the MCI is that most judges are experts and will therefore judge consistently with each other, while a judge whose pattern differs is inexpert. The MCI detects judges whose patterns of recommendations are inconsistent with those of the majority. In Jaeger's study of the MCI, eliminating judges whose patterns of ratings were aberrant resulted in little effect on one test (reading) but led to a substantially lower passing score for another test (mathematics).

It is important to recall, again, that while is customary to view interjudge consistency as desirable, it may not represent a legitimate criterion for evaluating the

standard-setting procedure if the judges were chosen to represent diverse perspectives (Cizek, 1996a). Jaeger (1988) himself notes that it seems more reasonable to eliminate the recommendations of some judges because they make poor absolute judgments rather than to eliminate their recommendations because their pattern of item recommendations is inconsistent relative to that of the majority."

When iterative procedures are used, it is also possible to show interjudge consistency by reporting round-to-round changes in variability of judgments, e.g. by showing correlation matrices showing the relationships among participants' judgments (Cizek, 1996a).

Reliability of the passing score. Subkoviak (1988) provides simple computational procedures for estimating two reliability indices from a single test administration. The agreement coefficient (p_0) represents the proportion of examinees consistently classified on two administrations of the mastery test (overall consistency). κ is a coefficient which represents the proportion of consistent classifications beyond the proportion that would be expected by chance. Subkoviak suggests acceptable reliability values for high-stakes and teacher-made tests.

Other reliability estimates have been recommended by Kane and Wilson (1984) and Jaeger (1989). Jaeger's index combines the standard error of measurement for the observed test scores and the standard error of the mean passing score recommended by the judges (see section on "Number of judges"). Kane and Wilson recommend using generalizability theory to obtain estimates for the variance components of judges and items, and if the data were collected in more than one standard-setting study, the variance component for occasions can also be estimated.

External Criteria

"Comparisons with external criteria tend to provide a check mainly on the "policy assumption", which claims that the standard is appropriate given the purpose of the decisions" (Kane, 1994). Kane discusses several external validity checks: criterion-related evidence (although this is hardly ever possible to implement in practice); comparisons to results of other standard-setting methods (a "second opinion" on the initial set of judgments, although it does not indicate which set is better), comparison to pass-fail decisions made with a different test, comparisons involving other assessment methods, comparisons of the group taking the test with the rate of competence in a similar population, and overall judgments by stakeholder groups of the pass rate.

As in other types of validity checks, none of these checks is decisive. "Presumably, the procedures used in standard setting were seen as the most reasonable way to set a passing score in that case. Therefore, the alternative sources of data that are readily available for external checks on validity are likely to be viewed as being at best comparable, and often inferior, to the procedures used to set the original passing score. As a result, a lack of correspondence between the original passing score and that suggested by the alternative approach is not compelling, and the evidence provided by the external validity checks tends to be highly ambiguous" (Kane, 1994).

Post hoc adjustments to cutoff scores

Even after a standard-setting panel has used carefully implemented procedures to arrive at a cutoff score, the score may be adjusted based on various criteria (Cizek, 1996a). Geisinger (1991) and Cizek (1996a) discuss sources of information that may be relevant in making adjustments; these are summarized in this section.

The first two types of information listed are traditionally employed and may be considered primary (Geisinger, 1991). The other types have generally been used indirectly.

Acceptable passing and failing rates

This information is explicitly considered in compromise models (Beuk, Hofstee, or deGruijter). Another use is by state licensing panels, who may, because of a desired percentage passing rate, set passing scores based on a given standard score from the distribution of test scores for a specific test administration. (Geisinger, 1991).

Relative costs of misclassification errors

The two types of errors are (a) passing someone who should not have passed and (b) failing someone who should have passed. Decision-theoretic models explicitly incorporate the estimated costs of each of these types of errors into the standard-setting process (e.g. Swaminathan et al., 1975).

Giving examinees the benefit of the doubt

Often these adjustments consist of raising or lowering the passing score by a fraction or multiple of the standard error of measurement of the test (Cizek, 1996a). Mehrens (1986) argued that the values underlying these considerations ought to be made explicit, and provides examples of the large effect such adjustments have on false positive and false negative decisions.

Difficulties with judges and their evaluations

Any of the issues discussed in the section on "factors affecting consistency of judgments" may constitute a reason to adjust the passing score. In these cases, the data of individual judges may have to be eliminated or another panel may have to be convened (Geisinger, 1991). Another possibility is to correct for rater effects, either for leniency or stringency or for judges' internal inconsistencies (Houston, Raymond & Svec, 1991). However, "it is an open question whether reducing variability is more defensible if attained through improved training, by more thorough group consensus building, or by statistical methods" (Cizek, 1996a).

Possibilities for retesting

Candidates should be permitted more than one chance to pass a test (Jaeger, 1988). However, when tests are given frequently, it seems reasonable to adhere stringently to the passing score. If the test is given infrequently, Geisinger (1991) suggests that the passing score might be adjusted in order to increase one's certainty that the candidate has indeed failed. Millman (1989) provides evidence that taking the test successive times increases the possibility of an incompetent candidate passing, and suggests specific formulas for increasing the passing score for repeat test-takers.

Societal or organizational needs

After the standard-setting process is completed, it may be found that the proposed passing point does not yield an adequate number of passers for organizational needs. It is then possible to compute what test score would yield an adequate number of passers. Then, using a normal curve table, it is possible to ascertain the probability that individuals scoring at that score would hold true scores at the minimally competent level. One could then decide whether to accept the lower cutoff score or to hold to the higher standard and recruit more examinees (Geisinger, 1991).

In the United States, it has been deemed unreasonable to consider individuals as passers of a test if in fact their test scores leave them little chance of being hired ("Uniform Guidelines on Employee Selection Procedures", 1978).

Adverse or disparate impact

The passing score may produce adverse impact on specific protected groups: sexual, racial or ethnic. In the US, a passing rate for a protected group which is less than 80% of the group with the highest rate generally be regarded as evidence of adverse impact. Geisinger (1991) describes how one may incorporate such

information into a modified Angoff standard-setting process: One Angoff panel sets a traditional passing score. A second panel sets a score representative of how they would expect an outstanding individual to score. The range between these scores represents the range within which a passing score might be set. Then, adverse impact ratios are calculated at each test score within this range and the score with the lowest disparate impact is employed as the final cut score.

Cizek's (1996a) closing remarks on the subject of adjustments to cutoff scores are that of course, the ideal approach is to avoid the problem through employing adequate standard-setting procedures in the first place: selection of judges, training, group monitoring, provision of appropriate data to judges, etc. Cizek recommends that a panel composed of informed experts be empowered to evaluate all the evidence and make qualitative decisions about adjusting cutoff scores. If post hoc adjustments are made, the rationale for using the particular adjustment must be clearly explicated and documented (Cizek, 1996a; Geisinger, 1991).

CHAPTER 4: STANDARD SETTING ON THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

The National Assessment of Educational Progress (NAEP) in the United States tests students in mathematics, reading, writing, science, history and geography, providing information to both educational professionals and to lay people about what students in American schools know and can do (Beaton & Zwick, 1992). NAEP issued its first results in 1970. Broad recognition of NAEP is indicated in its having become known since then as "The Nation's Report Card".

In the first assessments, NAEP results were reported in terms of *p*-values: the proportion of a group of students who could demonstrate a certain skill by answering a certain item correctly was reported (Bourque & Hambleton, 1993). However, it was desired to report the results in a way that would be more meaningful in terms of student performance. To this end, in 1982, the Educational Testing Service, the contractor then responsible for NAEP operations, introduced the "NAEP scale", in which performance of students in Grades 4, 8 and 12 was placed on a single, unidimensional, item response theory (IRT) scale (Lissitz & Bourque, 1995). Describing at what students at different points on the scale "know and can do" portrayed the increases in proficiency between the various points (Beaton & Allen, 1992), something which had not been possible when reporting only *p*-values.

Up to this point, NAEP only described student performance; NAEP did not compare students' actual performance to any description of what they ought to be able to do. However, in practice, NAEP results were given interpretations in terms of implicit standards (Linn & Dunbar, 1992). In 1988, this trend was formalized with the formation of the National Assessment Governing Board, given by NAEP the responsibility of identifying "appropriate achievement goals for each age and grade level in each subject area" (Koretz & Diebert, citing the NAEP Improvement Act). To achieve this goal, NAGB now sets performance standards, called *achievement levels,* on NAEP.

Both the ETS *scale anchoring* approach and the achievement level approach have been subject to controversy. In the next sections, both approaches will be described and then the controversial issues raised about each will be discussed.

Scale anchoring

Anchor levels were derived first by an empirical and then by a judgmental process. After obtaining the IRT scale for student performance in grades 4, 8 and 12, a few points on the scale, selected based on the distribution of scores, were described.

These were called *anchor points* or *anchor levels.* For example, the, anchor points on the NAEP scale were typically selected at 200, 250, 300 and 350, with 250 representing the mean of the scale and the other points representing standard deviation units. (The use of IRT in scale anchoring for NAEP is described in detail in Yamamoto & Mazzeo, 1992.) Note that this was a norm-referenced scale. Then, each item was examined empirically to see whether it discriminated between successive anchor points. The criteria for an item being an anchor point were: (a) at least 65% of examinees performing at or around the anchor level must answer the item correctly; (b) fewer than 50% of examinees performing at the next lower level should be able to answer the item correctly; (c) the difference between the two groups should be at least 30%, and (d) these *p*-values had to be based on a sample of at least 100 students (Lissitz & Bourque, 1995). The scale anchoring process is described in greater detail in Beaton and Allen, 1992.

Once the anchor points were identified, describing the content of each point was accomplished through a judgmental process. About twenty curriculum experts working in two independent groups were provided with *p*-values and other item information for the anchor items. They developed a description of each item that was based on common developmental themes among the items and which described their content in as much detail as possible. Differences in descriptions between the two groups were mediated by discussion and a final set of descriptions was agreed upon. This process resulted in a description of what students at each anchor point "know and can do", as described by their performance on a selected set of test items (Lissitz & Bourque, 1995).

Achievement levels

In response to the NAGB requirements, achievement levels began to be developed for NAEP, and were first used in 1990 (along with anchor levels). The Angoff procedure was chosen based on advice from standard-setting experts and Berk (1986), for nine reasons (Bourque & Hambleton, 1993) :

(1) It yields appropriate classification information.

(2) It is sensitive to student performance.

(3) It is sensitive to instruction and training.

(4) It is judged in the measurement literature to be statistically sound.

(5) It takes measurement error into account.

(6) It is easy to compute.

(7) It is generally easy to explain to lay people.

(8) It is generally credible to lay people.

(9) It can be applied to many item formats.

The modified Angoff procedure used by NAEP in 1990 resulted in three levels of achievement, known as *basic, proficient* and *advanced*, defined in terms of a percent-correct score on the NAEP assessment. In establishing these levels, experts were explicitly asked to respond to NAGB's policy position which called for defining what students should know and be able to do at grade levels 4, 8 and 12 (Lissitz & Bourque, 1995).

The procedure used to establish the achievement levels was as follows: After initial training, grade-level groups of judges were administered the NAEP assessment itself for that grade (about 60 items). Next, they received training in the modified Angoff methodology. The judges then rated the full item pool (around 200 items), estimating for each item the proportion of students who would be expected to answer the item correctly based on the policy definitions of each of the three levels. In the second iteration, judges again rated the items after having seen the *p*-values for each item from the 1990 NAEP administration. Thus, experts worked with each item, as distinct from the methodology used for anchor levels, where experts prepared descriptions from just the subset of items at each anchor level.

In the third iteration, the judges were provided with the aggregated results across items of the previous iterations, and this feedback was discussed in grade-level groups. They then completed a third rating. Then, items that distinguished between the achievement levels were identified and used in the preparation of written descriptions illustrating each of the three levels (Lissitz & Bourque, 1995) for each grade tested (Burstein et al., 1995).

Both the process used to establish the 1990 ratings and the meaningfulness of the final standards were subject to criticism. In response to the controversy, for the 1992 NAEP assessment, it was decided by American College Testing (ACT), the contractor responsible for implementing standard-setting procedures for NAEP, not to use the 1990 levels as a baseline but rather to undertake a new effort (Burstein et al., 1995). Technical improvements included replicable sampling procedures for selecting judges, comprehensive data analysis, and reliability and validity studies (ACT, 1992, 1993b).

In 1992, judges were given feedback on how their own ratings compared with other judges' ratings, so that outliers knew who they were and could consider this when they proceeded to later ratings. Also, the 1992 effort provided for a consideration

of intrajudge consistency, so that judges could improve their own consistency in later ratings (Lissitz & Bourque, 1995).

Another improvement was that in the 1990 mathematics assessment, judges were provided with generic definitions of the three achievement levels which they had to apply across grade levels and subject areas. However, in order to use generic definitions, judges need to operationalize them for the grade level and content with which they are working. The risk is that different judges operationalize the generic definitions differently, leading to variability in the standards. To avoid this problem, beginning in 1992, operational definitions were prepared before any item rating was done (Bourque & Hambleton, 1993). NAGB provided simple, policy-based definitions for each achievement level (basic, proficient, and advanced), from which panels of judges developed grade- and subject-matter-specific descriptions of the levels (Burstein et al., 1995). The policy-based definitions were as follows:

> *Basic.* This level, below proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at Grades 4, 8 and 12. For 12th grade, this is higher than minimum competency skills (which normally are taught in elementary and junior high schools) and covers significant elements of standard high-school-level work.
>
> *Proficient.* This central level represents solid academic performance for each grade tested – 4, 8, and 12. It reflects a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At Grade 12, the Proficient level encompasses a body of subject-matter knowledge and analytic skills and of cultural literacy and insight that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.
>
> *Advanced.* This higher level signifies superior performance beyond proficient grade-level mastery at Grades 4, 8 and 12. For 12th grade, the advanced level shows readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement exams (Phillips et al., 1993).

In the case of the 1992 mathematics assessment, the judges used these definitions, the NAEP mathematics framework, and their experience with NAEP assessments to arrive at content-based descriptions of each achievement level. These

descriptions were then used by panels of judges (teachers, other educators and noneducators) in reviewing the NAEP items. The judges estimated the percentage of students at the borderline of each achievement level who would respond correctly to the items. The average judgments on a final set of ratings were mapped onto the NAEP scale (More detail about the rating process is provided in ACT, 1993a) (Burstein et al., 1995). For future efforts the definitions were to be prepared even before item development (Bourque & Hambleton, 1993).

The 1992 process established nine scale points corresponding to the minimal scores for each of the three achievement levels at each grade level. NAGB adjusted the levels to one standard error below these scale points (Burstein et al., 1995).

Panelists also made final refinements of the descriptions of the achievement levels. Exemplar items which met specific empirical criteria were also chosen for each level and grade. Exemplar items were selected by a panel based on the quality of the item, the coverage of content for the set of exemplars as a whole, and the grade appropriateness for items that were used at more than one grade. The empirical criteria for the Basic level were, for example: (a) the expected p-value for students at the cut point for the Basic level were greater than .51; (b) the content of the item matched the content of the operationalized description of Basic; and (c) the empirical *p*- value for the item had to be higher than the empirical *p*-value for items selected as exemplars for the Proficient level (Bourque, 1993, pp. 9-10).

Validity on NAEP

Unfortunately, most of the material written by NAGB describing validity issues for NAEP (including exact procedures followed in training judges and rating items, and efforts and measurements made for internal and external validity consistency ratings) is unpublished. The admittedly sparse material in this section is gleaned from references to published material.

Internal validity

Shepard et al. (1993) employed a number of internal validity checks on NAEP results, including comparisons between the standards set on different types of items (e.g. multiple-choice versus extended-response) and the standards set in different areas of content. These analyses indicated that the results were consistent across content areas but that the standards based on extended response items tended to be significantly higher than those based on multiple-choice items. The reasons for the observed

differences could not be determined from the data in that study, but the internal checks were effective in revealing the existence of a potential problem (Shepard et al., 1993).

External validity

NAEP achievement levels were compared to teachers' judgments about student proficiency, to other indicators of student performance, and to results of international assessments (Shepard et al., 1993).

Measuring consistency

An approach similar to van der Linden's was used to measure intrajudge consistency in the achievement-level setting procedures (Cizek, 1996a).

Concerns about NAEP standard-setting methods

First, the concerns about scale anchoring for NAEP tests will be listed, followed by the concerns surrounding achievement levels. Then, several criticisms applying to both methods will be discussed.

Criticisms of setting anchor levels on NAEP.

1) There is a problem with content validity because a very small number of items are used to describe the anchor points; large numbers of items are omitted from consideration because they do not anchor at any of the selected points (Lissitz & Bourque, 1995).

2) The use of IRT methodology is not justified, because the domain being studied was ill-defined; an ill-defined domain could not possibly be based on an underlying unidimensional trait as required by IRT (Forsyth, 1991). The response of ETS to this criticism was that the fact that the domain produced discriminating items (anchor points) proved that it was ordered (Beaton & Johnson, 1992). However, the point remained controversial.

Criticisms of reporting of anchor levels.

3) The early anchor-based reports included "grade-based" descriptions describing when the material in the anchor item is taught rather than the grade in which it is typically mastered by students. For example, the panel characterized Level 200 in part as "material typically covered by the third grade" (Mullis, Dossey & Owen et al., 1991, pp. 6-7). Such descriptions were highly controversial and were eliminated entirely by 1992.

4) Anchor report descriptions in terms of predictions of students' likelihood of success in later activities were criticized because of a lack of any validating evidence about the actual performance of students in later activities (e.g. Forsyth, 1991).

Criticisms of achievement levels.

Burstein et al. (1995) investigated the validity of the descriptions and exemplar items used in the 1992 NAEP mathematics assessment, in which achievement levels were used. The researchers found five main problems in the achievement level design, analyses and reporting:

1) The item pool provided inadequate or no coverage of some content attributes used in the descriptions of the achievement levels, for some of the achievement levels. Thus it is impossible to say whether students who scored at the levels in question actually could perform tasks implied by the descriptions.

2) "The definitions of the achievement levels overlap considerably and frequently differ in minor or unclear ways." The authors found the levels ambiguous and predict that other experts, and certainly lay people, will too.

3) Many of the items selected as exemplars of the achievement levels were misleading in terms of actual performance of students. In some cases, fewer than half of the students performing within the range of a given achievement level answered the exemplar item correctly, and, in other cases, more than 75% of the students performing at a given level answered correctly an item intended to be an exemplar of the next highest level.

4) Students scoring at a given level frequently showed what many people might consider to be unreasonably low rates of success on some groups of items linked to descriptions at their level. For example, in some instances, for half the items in a given set, fewer than half of the students at that level answered correctly.

5) Items that actually did differentiate among achievement levels in terms of actual student performance did not correspond well to the achievement-levels descriptions.

Due to what they believed to be the severity of these problems, Burstein et al. recommended that the 1992 descriptors and exemplars not be used in reporting NAEP results. NAGB did not alter the 1992 reporting in mathematics, but it did alter the selection of exemplars in its next effort (reading) in response to the findings.

Several concerns with respect to reporting apply to both anchor and achievement levels (Koretz & Deibert, 1995):

1) In many press reports about NAEP, even those by noted educators, an unexpected misinterpretation took place: the percentage of students who score above an anchor point (or at an achievement level was interpreted to mean the percentage of

students who answer an anchor (or exemplar) item correctly. This misinterpretation came about from assuming that only students scoring above the anchor point answered the item correctly, whereas actually some of the students scoring below the anchor point answered the item correctly (Linn & Dunbar, 1992).

2) An unintended effect of both the anchor point and achievement level methods was to encourage the misunderstanding of performance as discontinuous. The proper interpretation of an anchor or achievement level is a point on a continuum at which the rate of success on tasks reaches a certain level, but the levels were interpreted instead to mean the point at which students could do the things noted in the anchor-point descriptions. NAEP experience suggests that reporting *p*-values along with standards-based reporting is not enough of an antidote to this problem, since few reporters used *p*-values, and even fewer used them correctly.

3) The NAGB report did not explain the implications of the judgmental nature of the achievement levels: namely, that the resulting standards depend on the judges selected and the methods used. Most of the press reports made no reference at all to the judgmental nature of the achievement levels.

4) The press articles showed a tendency to rely on familiar metrics or even to convert novel metrics into familiar ones. One example of this was conversion of anchor points into predictive expectations for specific grades.

Several problems arose with respect to the judgmental methods used on NAEP:

1) *Interrater consistency.* For the 1990 mathematics assessment, NAGB decided that both educators and non-educators should participate on judgment panels. Non-educators included professionals who used mathematics such as engineers, developers of children's television programming, and individuals representing national stakeholder groups who had a background in mathematics. The composition of the panel caused several problems. The stakeholder groups came with their own agenda and distracted attention from the task at hand, standard-setting. The noneducators lacked general background knowledge about NAEP, large-scale assessment, and, in some cases, lacked knowledge about mathematics as currently conceptualized in educational circles. This diversity led to a lack of consensus in the item ratings and resulting standards.

Interestingly, the standard setting process was repeated in four states with panels consisting mostly of teacher educators, but even this did not lead to more homogeneous ratings (Bourque & Hambleton, 1993). The following point has already been mentioned in several contexts in this review: Variability among raters, while a concern, is not necessarily a problem: with such broad representation, why would one expect or desire high levels of consistency? (Lissitz & Bourque, 1995).

2) *Validity of the modified Angoff procedure.* The recent challenge to the fundamentals of all item-judgment (test-centered) procedures was made in response to NAEP, and was mentioned in Chapter 2. "The purported ease of implementation–indeed, the validity–of the Angoff method has... recently been challenged... A report of the National Academy of Education studied implementation of a modified Angoff approach used to set standards for 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 candidates, may be impossible for participants to acquire or 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 et al., 1993, p. xxiv).

CHAPTER 5: MEDICAL CERTIFICATION TESTS

Until 1981, the National Board of Medical Examiners (NBME) in the United States used norm-referenced standards on its Part I and Part II Examinations. Part I is normally taken after the first two years of medical school, and Part II upon graduation (after four years of school). Beginning in 1981, standards for Part I and Part II were based on the performance of a criterion group defined to include reference group examinees from the preceding four years. Although this resulted in more stable standards than previously, the standards still shifted whenever the performance of the reference group changed. Also, the standard-setting procedure was still based on the performance of a group of examinees rather than a specified level of mastery of content, while the latter was considered more intuitively satisfying for a licensure examination. Therefore, in conjunction with the introduction of the new comprehensive Part I and Part II tests in 1991, NBME adopted a new standard-setting plan. The following sections describes the research behind the plan and the tentative plan that was adopted in a three-phase process. (Nungester et al., 1991).

During Phase 1, NBME investigated the sensibility, psychometric characteristics, and acceptability of several content-based standard-setting procedures for the comprehensive Part I examination, with medical-school faculty as subjects. Two techniques, Angoff and Ebel, were studied, crossed with three methods. In Method I, judgments were made without any information about examinee performance. In Method II, judges provided initial judgments without performance information (these judgments were the basis for the Method I results); then, statistical information was provided and judges were allowed to revise their initial judgments. For Method III, judges were provided with performance data from the outset. The procedure used is described in detail in Swanson, Dillon and Ross (1990).

"Among all the six procedures, the Angoff Method II resulted in by far the most reproducible estimate of the passing score....There was a strong consensus among the judges that Method II was the most appropriate. Method II allowed the standard-setting process to be influenced initially by judges' experience in medical education and content expertise, and subsequently by examinee performance. Both of these components were viewed as essential for the resulting standards to be informed and meaningful. Reactions to the Angoff and Ebel techniques were mixed. Those favoring the Angoff technique found it simple and straightforward to use. Those favoring the Ebel technique thought that item relevance should affect the standard-setting judgments, though some felt that relevance should be addressed during test

development, not standard-setting." (Swanson, Dillon & Ross, 1990). The Angoff Method II approach was selected.

"During Phase 2, the procedure from Phase 1 was refined and its use was extended to the comprehensive Part II examination... [As of the writing of Nungester at al. (1991)], Phase 3 was still under way. In Phase 3, small groups of faculty judges from several dozen medical schools reviewed items to appear on the initial comprehensive examinations. Participating faculty members were nominated because of their content expertise and their broad familiarity with medical education, medical students, and the purpose of the examinations. For the comprehensive Part I examination, both basic science and faculty participated; the majority were directors of basic science courses or required clinical clerkships. Course and clerkship directors also participated in standard-setting groups for the comprehensive Part II examination; directors of primary care residency programs were also included. Overall, participating faculty came from a broad cross-section of medical schools, and women and minorities were well represented" (Nungester et al., 1991).

The content-based standard-setting procedure is typical of a modified Angoff procedure. The steps are as follows:

"1. Provision of background information to judges. The information provided includes an introduction to (a) the purpose, format, and content of the comprehensive examinations; (b) the general area of standard-setting; and (c) the modified Angoff approach used to set standards.

2. Discussion of the concept of a "borderline examinee". Based on their experience in working with medical students, judges are asked to discuss the characteristics of borderline examinees (students who should receive the lowest passing score on the examination).

3. Initial review of test items. After having some practice with the procedure, judges are asked to independently predict the percentage of borderline examinees that would answer each item correctly.

4. Reconsideration of initial judgments. Based upon projections of examinee performance and the test material overall, each judge next reviews the initial judgments. If there is a perceived discrepancy between the projections and the judge's expectations, then the judge is encouraged to reevaluate the original decision and, if warranted, may revise his or her prediction of the percentage of borderline examinees that would answer each item correctly.

5. Calculation of the pass-fail standard. The standard is calculated by averaging the revised judgments from Step 4 across judges and items. The pass-fail

standard is provided to the Comprehensive Committee charged with the responsibility for standard-setting." (Nungester et al., 1991).

A general objection to setting passing scores is that if competence is a continuous variable, why penalize someone who scores just below the cut score? For the case of medical licensing tests, such a procedure is justified because even borderline passing candidates will select not just incorrect choices, but a "disturbing number" of actually dangerous ones (Juul & Loewy, 1988).

Equating of test forms of the NBME will be employed to avoid shifts in standards over time. In addition, a systematic annual review will monitor shifts and trends in fail rates. The Committee will also undertake a triennial review of standard-setting procedures, potentially leading to the determination of a new standard or new standard-setting procedures (Nungester et al, 1991).

It has been suggested that in the future, "sampled constituencies be expanded to include additional groups that are knowledgeable and informed about the NBME examination standards. Random samples of medical practitioners, chief administrators of teaching hospitals, nurses, members of hospital staffs, or informed members of the general public may provide further insights regarding examination standards" (Orr & Nungester, 1991).

Repeated test-taking. Allowing unsuccessful candidates to repeat the test increases the risk of false positives, because each test is subject to measurement error (Millman, 1989). It is crucial to minimize incidence of false positives in medical licensing tests, because of the grave risks of passing incompetent candidates. On the other hand, licensure boards must avoid mistakes against individuals because this can lead to litigation.

Millman (1989) makes several suggestions for correcting the problem of false positives over repeated testing: additional testing for candidates scoring near the borderline, an indifference zone within which competent candidates have an acceptably low chance of being misclassified, or, if these methods are unfeasible, more stringent passing requirements for subsequent attempts.

Millman discusses several variations on raising the passing score for subsequent attempts, including averaging the scores of all attempts and averaging the scores of the two most recent attempts. In Millman's hypothetical example, the method which provided the least number of false classifications was increasing the passing score as a function of the number of attempts.

Performance assessment. Standard-setting on medical certification tests that consist of performance assessments is discussed in several articles: Clauser, Clyman,

Margolis and Ross (1996); Rothman et al., 1991; and Rothman, Cohen and Ross, 1990.

SUMMARY AND RECOMMENDATIONS

Methods

Two main types of standard-setting methods exist: test-centered models, which involve judgments about test items, and examinee-centered methods, which involve judgments about examinees. The modified Angoff method, a variation on one of the test-centered models that includes provision to judges of normative data such as empirical item difficulties, has seen the most use recently of all the methods, both in educational assessments and medical certification testing.

The choice of passing score, ultimately, is fairly arbitrary; the best we can do is to show that the passing score and its associated performance standard are not unreasonable. Recently, more attention has been paid in the literature to the issue of ensuring that a standard-setting application is valid (Kane, 1994), including documentation of each stage of the particular standard-setting procedure (Cizek, 1996a).

Most recently, research based on the field of judgment and decision-making has questioned the validity of judgments made by judges in test-centered standard-setting methods (Mitzel, Lewis & Ross, 1996).

Educational tests

The educational community, following the lead of NAEP, apparently has settled on the modified Angoff procedure involving use of empirical data in the second round, as the best of the available standard-setting methods. The modified Angoff procedure is easy to compute, credible, has been judged to be statistically sound, and [has been thought to] yield appropriate classification information (Bourque & Hambleton, 1993).

One approach to assessing the quality of NAEP's current standard-setting methods is that although there were problems in the past, by now they have "worked out all the kinks" and their current standard-setting procedures are acceptable. After each NAEP administration, NAEP has made efforts to rectify the problems that were raised with its standard-setting methods and reporting, if not for the assessment under criticism then for the next one. The problems raised by critics of the 1992 NAEP assessments with respect to reporting could be rectified through more stringent guidelines for selection of exemplar items and improved explanations to the public of the standard-setting process.

A second approach is that there remain serious problems with standard-setting procedures on NAEP, and that these problems may not be rectifiable within the

framework of a modified Angoff procedure. For the 1992 NAEP, reported problems included (Burstein et al., 1995) inadequate coverage of content, unclear and overlapping definitions of achievement levels, and about items that differentiated among achievement levels in terms of actual performance not corresponding to the achievement-level descriptions. The doubts raised by Shepard at al. (1993) and Mitzel, Lewis & Ross (1996) speak to the same underlying problem as is probably the cause of these specific problems found by Burstein et al.: difficulty in constructing and maintaining stable definitions of basic constructs such as minimal competence for each proficiency level.

The field seems to be narrowed down to the modified Angoff approach and the "Bookmark Approach" of Mitzel, Lewis and Ross (1996). In the "Bookmark Approach", judges set cut points in test booklets in which the items are presented in order of IRT ability scale "location". The following paragraphs evaluate the relative advantages and disadvantages of the Bookmark Approach compared to the more familiar modified Angoff approach.

Mitzel, Lewis and Ross (1996) suggest, based on theories of judgement and decision making, that the Bookmark Approach is more valid than the popular modified Angoff approach. The Bookmark Approach has been used so far in one standard-setting procedure. Clearly, the Bookmark Approach has not been subjected to the extensive usage that the modified Angoff procedure has been through; further use of the Bookmark Approach could uncover problems. For example, it could be argued that the judges used in the Bookmark Approach of Mitzel et al. were not representative group: the group consisted of (a) four teachers, one from each of four regions of the United States, (b) one content expert selected based on a national reputation for expertise in the content area, and (c) one CTB content area editor. Also, the standard-setting procedure performed using the Bookmark Approach has not yet been subject to objective review.

On the positive side, the Bookmark Approach has thus far appeared to be a model of clarity compared to other standard-setting procedures, including some of NAEP's. For example, in other standard-setting studies, the problem existed that judges did not understand the standard-setting process or admitted lack of confidence in their ratings, while for the Bookmark Approach the judges were reported to be enthusiastic and confident.

One difference between the modified Angoff procedures and the Bookmark Approach is that in the Bookmark Approach, the rank order of item difficulty is known to raters at the outset. One of the advantages of the modified Angoff procedure is that the judges rate the items "purely", and only later take into account normative data such

as difficulty ratings. However, it may be that the form in which the difficulty ratings are provided in the Bookmark Approach, ordering of the items from easiest to most difficult, without *p*-values, permits judges to focus on item content, while at the same time considering item difficulty only as secondary data (similar to the idea of providing normative data in the second round rather than initially). It could be argued that provision of item data in this form organizes the judges' task; rather than having to simultaneously consider for each item the two dimensions of difficulty and content, they can focus just on a less confusing task, considering the one dimension of item content.

Based on this review, reasonable recommendations seem to be to use either (a) a modified Angoff approach, with careful attention to all aspects of the standard-setting process, or (b) be a pioneer of the Bookmark Approach. A third possibility is to use both methods and compare the results. Similar results from both methods would point to the validity of the standards set; differing results would require interpretation at that point. If the Bookmark Approach is selected, it should be described in detail, since it is a new method.

Medical tests

The use of the modified Angoff approach on medical certification tests has been less controversial than its use on NAEP-like tests. The reason for this may be that for medical tests, the definition of competence is not under dispute (cf. Kane, 1994) and the content domain is well-defined. Also, the domain of possible expert judges is somewhat more limited and more uniform.

REFERENCES

American College Testing. (1992). *Design document for setting achievement levels on the 1992 NAEP in mathematics, reading, and writing.* Iowa City, IA: Author.

American College Testing (1993a). *Description of mathematics achievement level setting process and proposed achievement levels descriptions.* Washington, DC: National Assessment Governing Board.

American College Testing (1993b). *Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading,and Writing: A technical report on reliability and validity.* Iowa City, IA: Author.

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.

Applied Measurement in Education, 8, (1995), special issue entitled "Standard Setting for Complex Performance Tasks".

Beaton, A. E. & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17,* 191-204.

Beaton, A. E. & Johnson, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17,* 191-204.

Beaton, A. E. & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics, 17,* 95-109.

Berk, R. A. (1976). Determination of optional cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45,* 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.

Berk, R. A. (1995). Something old, something new, something borrowed, a lot to do. *Applied Measurement in Education, 8,* 99-109.

Berk, R. A. (1996). Standard setting: The next generation (Where few psychometricians have gone before!). *Applied Measurement in Education, 9,* 215-235.

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.

Bourque, M. L. & Hambleton, R. K. (1993). *Measurement and Evaluation in Counseling and Development, 26,* 41-47.

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.

Burstein, L., Koretz, D., Linn, R., Sugure, B., Novak, J., Baker, E. L., & Harris, E. L. (1995/1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance.*Educational Assessment 3,* 9-51.

Busch, J. C. & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27,* 145-163.

Chang, L., Dziuban, C. D., Hynes, M. C. & Olson, A. H. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education, 9,* 161-173.

Chi, M., Glaser, R., & Farr, M. (Eds.), (1988). *The nature of expertise.* Hillsdale. NJ: Lawrence Erlbaum.

Chi, M., Glaser, R., & Rees, F. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence,* vol. I (pp. 17-76). Hillsdale, NJ: Lawrence Erlbaum.

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

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

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

Clauser, B. E., Clyman, S. G., Margolis, M. J., & Ross, L. P. (1996). Are fully compensatory models appropriate for setting standards on performance assessments of clinical skills? *Academic Medicine, January Supplement, 71,* S90-S92.

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. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22,* 263-269.

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

Educational Testing Service. (1984). *NTE programs test analysis booklet–core battery.* Unpublished statistical report SR-84-19. Princeton, NJ: Author.

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

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38290-38315.

Fehrmann, M. L., Woehr, D. J., & Arthur, W. (1991). The Angoff cutoff score method: The impact of frame-of-reference rater training. *Educational and Psychological Measurement, 51,* 857-872.

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

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice, 10,* 3-9, 16.

Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. *Educational Measurement: Issues and Practice, 10,* 17-22.

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

Gross, L. J. (1985). Setting cutoff scores on credentialing examinations. *Evaluation and the Health Professions, 8,* 469-483.

Hambleton, R. K.& Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences.* (pp. 367-396). Berkeley: CA: McCutchan.

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

Hobson, C. J. & Gibson, F. W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. *Academy of Management Review, 8,* 640-649.

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.

Houston, W. M., Raymond, M. R., & Svec, J. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15,* 409-421.

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. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. *Applied Measurement in Education, 1,* 17-31.

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,* 3-6, 10, 14.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8,* 15-40.

Juul, D. & Loewy, E. H. (1988, April). *Setting and using cutoff scores on tests used for certification.* Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA.

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

Kane, M. T. & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Journal of Educational Measurement, 24,* 56-64.

Koretz, D. & Deibert, E. (1995/1996). Setting standards and interpreting achievement: A cautionary tale from the National Assessment of Educational Progress. *Educational Assessment, 3,* 53-81.

Langsley, D. G. (1987). Prior ABMS conferences on recertification. In J. S. Loyd & D. G. Langsley (Eds.), *Recertification for medical specialists* (pp. 11-30). Evanston, IL: American Board of Medical Specialties.

Linn, R. L. & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29,* 177-194.

Lissitz, R. W. & Bourque, M. L. (1995). Reporting NAEP results using standards. *Educational Measurement: Issues and Practice, 2,* 14-23, 31.

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,* 121-141.

Mehrens, W. A. (1986). Measurement specialists: Motive to achieve or motive to avoid failure? *Educational Measurement: Issues and Practice, 5,* 5-10.

Melican, G. J., Mills, C. N., & Plake, B. S. (April, 1987). *Accuracy of item performance predictions based on the Nedelsky standard setting method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Melican, G. J. & Thomas, N. (1984, April). Identification of items that are hard to rate accurately using Angoff's standard setting method. Paper presented at the

annual meeting of the American Educational Research Association, New Orleans.

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. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10,* 9-20.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18,* 5-9.

Mills, C. N. & Melican, G. J. (April, 1987). *An investigation of three methods for adjusting cutoff scores.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice, 10,* 7-10.

Mitzel, H. C., Lewis, D. M., & Green, D. R. (1996). *Standard setting: A bookmark approach.* Paper presented at the CCSSO National Conference on Large Scale Assessment.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *NAEP 1992 mathematics report card for the nation and the states* (Report No. 23-ST02). Washington, DC: U. S. Department of Education, National Center for Education Statistics.

National Assessment of Educational Progress Improvement Act of 1988, Pub. L. No. 100-297, §3403.

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

Norcini, J., Shea, J., & Grosso, L. (1991). The effect of numbers of experts and common items on cutting score equivalents based on expert judgment. *Applied Psychological Measurement, 15* 241-246.

Nungester, R. J., Dillon, G. F., Swanson, D. B., Orr, N. A., & Powell, R. D. (1991). Standard-setting plans for the NBME Comprehensive Part I and Part II Examinations. *Academic Medicine, 66,* 429-433.

Orr, N. A. & Nungester, R. J. (1991). Assessment of constituency opinion about NBME examination standards. *Academic Medicine 66,* 465-470.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., & Williams, P. L. (1993). *Interpreting NAEP scales.* Washington, DC: National Center for Education Statistics.

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., & 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,* 57-83.

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

Rothman, A. I., Cohen, R., Dirks, F. R., Poldre, P., & Ross, J. (1991). Validity and reliability of a domain-referenced test of clinical competence for foreign medical graduates. *Academic Medicine, 66,* 423-425.

Rothman, A. I., Cohen, R., and Ross, J. (1990). Evaluating the clinical skills of foreign medical school graduates participating in an internship preparation program. *Academic Medicine, 65,* 391-395.

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 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, J. K., Smith, R. L., Richards, C., & Barnhardt, S. (1988, March). *The optimal number of judges to use in setting passing scores.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indexes for mastery tests. *Journal of Educational Measurement, 25,* 47-55.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement, 12,* 87-98.

Swanson, D. B., Dillon, G. F., and Ross, L. E. P. (1990). Setting content-based standards for National Board exams: Initial research for the comprehensive Part I examination. *Academic Medicine, September Supplement, 65,* S17-S18.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge consistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295-308.

Yamamoto, K. & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17,* 153-173.

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