An NCME Instructional Module on

Setting Performance Standards: Contemporary Methods

Gregory J. Cizek, *University of North Carolina-Chapel Hill*
Michael B. Bunch, *Measurement Incorporated*
Heather Koons, *North Carolina Department of Public Instruction, Raleigh*

This module describes some common standard-setting procedures used to derive performance levels for achievement tests in education, licensure, and certification. Upon completing the module, readers will be able to: describe what standard setting is; understand why standard setting is necessary; recognize some of the purposes of standard setting; calculate cut scores using various methods; and identify elements to be considered when evaluating standard-setting procedures. A self-test and annotated bibliography are provided at the end of the module. Teaching aids to accompany the module are available through NCME.

Keywords: *cut scores, performance standards, standard setting*

Fewer than 10 years have elapsed since the publication of the first Instructional Topics in Educational Measurement Series (ITEMS) module on standard setting in *Educational Measurement: Issues and Practice* (Cizek, 1996a). Nevertheless, since that time, a great deal of research, reconceptualization, and refinements to the methods of standard setting have transpired. In the earlier module, common standard-setting procedures—primarily applicable to selected-response format testing—were described, including the Contrasting Groups and Borderline Groups methods (Livingston & Zieky, 1982), and the Angoff (1971), Ebel (1972), and Nedelsky (1954) methods. So-called "compromise" methods by Beuk (1984) and Hofstee (1983) were also described.

While many of the aforementioned methods remain defensible routes for setting performance standards, other methods have been introduced. These contemporary methods have provided viable options for addressing evolving standard-setting controversies and challenges. For example, one goal of some new methods has been to reduce the cognitive burden placed on participants[1] to form and consistently apply conceptualizations of a hypothetical minimally qualified examinee in making judgments about probable success on individual test items. Another goal of emerging methods has been to provide a satisfactory way to establish standards on performance tests, that is, on tests that do not consist of dichotomously scored items, but contain polytomously scored samples of examinee work. As the consequences and costs of standard setting have escalated, research in the area of standard setting has attempted to derive methods that are more intuitive to participants and stakeholders and which can be implemented efficiently.

In addition to these changes, the standard-setting landscape has changed in other fundamental ways. A few examples of these profound changes are described.

Standards-Referenced Testing

Traditional ways of thinking about tests as yielding either norm- or criterion-referenced interpretations became outdated with the introduction of *standards-referenced* testing. Traditional standard-setting methods were developed largely for contexts in which only two categories (e.g., pass/fail) were required. The introduction of standards-referenced testing was accompanied by increased interest in defining more than two categories or performance levels. A prominent national testing program, the National Assessment of Educational Progress (NAEP), was one of the first, highly visible testing programs to express performance according to a graded series of performance levels: *Basic*, *Proficient*, and *Advanced*.

New *Standards for Educational and Psychological Testing*

In 1999, the three sponsoring entities for the *Standards*—the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education—issued revised standards for sound testing practice. This edition of the *Standards* highlights the importance of setting performance standards.

Gregory J. Cizek is Professor of Educational Measurement and Evaluation, School of Education, CB3500, University of North Carolina, Chapel Hill, NC 27599-3500; cizek@unc.edu. His areas of specialization are standard setting and testing policy.

Michael B. Bunch is Vice-President, Measurement Incorporated, Durham, NC. His areas of specialization are large-scale assessment design and standard setting.

Heather Koons is Project Manager, North Carolina Department of Public Instruction. Her areas of specialization are reading and science test development.

For testing in general, the *Standards* note that:

> A critical step in the development and use of some tests is to establish one or more cut points dividing the score range to partition the distribution of scores into categories. . . . [C]ut scores embody the rules according to which tests are used or interpreted. Thus, in some situations the validity of test interpretations may hinge on the cut scores. (p. 53)

And, in the specific case of licensure and certification tests,

> The validity of the inferences drawn from the test depends on whether the standard for passing makes a valid distinction between adequate and inadequate performance. (p. 157)

The 1999 version also includes new guidelines for standard setting. Among the guidance in the new *Standards* are:

> *Standard 1.7:* When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth. (p. 19)
>
> *Standard 2.14:* Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (p. 35)
>
> *Standard 2.15:* When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same or alternate forms of the instrument. (p. 35)
>
> *Standard 4.19:* When proposed interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented. (p. 59)
>
> *Standard 4.20:* When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria. (p. 60)
>
> *Standard 4.21:* When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way. (p. 60)
>
> *Standard 6.5:* When relevant for test interpretation, test documents should . . . include item level information, cut scores, . . . (p. 69)
>
> *Standard 14.17:* The level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test. (p. 162)

Federal Legislation

At the national level, at least two wide-ranging laws have affected the practice of standard setting. The Individuals with Disabilities Education Act (1997) requires greatly expanded participation of students with special needs in large-scale assessment programs. Among other regulations, the Act requires states to: (1) include children with disabilities in general state and district-level assessment programs; (2) develop and conduct alternate assessments for students who cannot participate in the general programs; and (3) provide public reports on the performance of special needs students with the same frequency and detail as reports on the assessment of nondisabled children. Developing new approaches to establishing performance standards for the required alternate assessments, which often comprise novel or nontraditional formats, has proven to be a significant standard-setting challenge.

A second piece of far-reaching legislation was enacted in 2001. The No Child Left Behind (NCLB) Act (2001) requires states to: (1) develop challenging content standards in reading, mathematics, and science; (2) develop and administer assessments aligned to those standards; and (3) (of particular relevance to standard setting) establish three levels of high achievement (*Basic, Proficient,* and *Advanced*) to describe varying levels of mastery of the content standards.

These three phenomena taken together—the rise of standards-referenced testing, the publication of new *Standards for Educational and Psychological Testing,* and recent federal legislation—have necessitated greater attention to standard setting than perhaps ever before. Much has been demanded of the technology of standard setting. New methods have been developed to meet new contexts and challenges, and substantially greater scrutiny and awareness of standard setting by policymakers, educators, and the public have resulted. This module is an attempt to catch up on these fast-paced changes.

In the following sections, we provide an update on concepts and methods of setting performance standards. First, we describe what is meant by *standard setting* and we provide a rationale for the need for setting standards. Next, we list some general considerations that warrant attention in any standard-setting procedure. Then, we describe three specific methods, introduced since the publication of the earlier module, which have found fairly wide usage in achievement testing contexts. These methods are presented in how-to format which, it is hoped, will provide sufficient detail to actually enable readers to use the method to obtain cut scores in a relevant situation. The final section of the module presents guidelines for evaluating standard setting. An annotated bibliography and self-test appear at the end of this module.

Definition of Standard Setting

It might seem obvious that what is called *standard setting* is the process by which a standard or cut score is established. In reality, however, standard setting is not so straightforward. For example, participants in a standard-setting process rarely *set* standards; rather, a standard-setting panel usually makes a recommendation to a body with the actual authority to implement, adjust, or reject the standard-setting panel's recommendation (e.g., state board of education, medical board, licensing agency).

It is now a widely accepted tenet of measurement theory that the work of standard-setting panels is not to search for a knowable boundary between categories that exist. Instead, standard-setting procedures enable participants to bring to bear their judgments in such a way as to *translate* policy decisions (often, as operationalized in performance level descriptors) into locations on a score scale; it is these translations that create the effective performance categories. This translation and creation are seldom, if ever, purely statistical, im-

partial, apolitical, or ideologically neutral activities. As noted in the *Standards for Educational and Psychological Testing,* standard setting "embod[ies] value judgments as well as technical and empirical considerations" (AERA/APA/NCME, 1999, p. 54). From this perspective, it is clear that what psychometrics as a social science can contribute to the practice of standard setting is as much social as it is science. As Cizek (2001b) has observed: "Standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other" (p. 5). Nonetheless, psychometricians have developed and continue to refine methods for negotiating these currents, and for aiding participants in bringing their judgments to bear in ways that are reproducible, informed by relevant sources of evidence, and fundamentally fair to those affected by the process.

One definition of standard setting, suggested by Cizek (1993), highlights the procedural aspect of standard setting and draws on both legal theory of due process[2] and traditional definitions of measurement. According to Cizek, standard setting is "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (p. 100).

Kane (1994) has provided a definition of standard setting that highlights the conceptual nature of the endeavor. According to Kane:

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

Finally, two additional observations are warranted. Despite Kane's (1994) attempted clarification, the term *performance standard* is frequently used as a synonym for the terms *cut score, achievement level,* or *passing score.* It is equally important to recognize that important decisions rest on two different kinds of standards that combine to make interpretation of test results meaningful; these are often referred to as *content standards* and *performance standards. Content standards* is the term used to refer to statements that describe specific knowledge or skills over which examinees are expected to have mastery for a given age, grade level, or field of study. Whereas content standards delineate the referent (i.e., the "what") of testing, performance standards define "how much" or "how well" examinees are expected to perform in order to be described as falling in a given category.

Need for Standard Setting

A fundamental issue in standard setting is the purpose for setting standards in the first place. From one perspective, the general need for standard setting is clear: Decisions must be made. As stated elsewhere:

> There is simply no way to escape making decisions. . . . These decisions, by definition, create categories. If, for example, some students graduate from high school and others do not, a categorical decision has been made, even if a graduation test was not used. (The decisions were, presumably, made on *some* basis.) High school music teachers make decisions such as who should be first chair for the clarinets. College faculties make decisions to tenure (or not) their colleagues. We embrace decision making regarding who should be licensed to practice medicine. All of these kinds of decisions are unavoidable; each should be based on sound information; and the information should be combined in some deliberate, considered, defensible manner. (Cizek, 2001a, p. 21; see also Mehrens & Cizek, 2001, pp. 478–479)

Certainly, decisions can be made on information other than, or in addition to, that yielded by tests. Indeed, the *Standards for Educational and Psychological Testing* state that "a decision or characterization that will have major impact on a student should not be made on the basis of a single test score" (AERA/APA/NCME, 1999, p. 146). In one sense, of course, this recommendation is always heeded. For example, a single measure such as the SAT for college admissions should be used with other criteria (e.g., high school graduation, grade point average, and so on). On the other hand, the information yielded by tests routinely figures prominently into decisions such as placement in a remedial or gifted program, selection of employees, awarding of scholarships, licensure to practice in a profession, and others. This is perhaps the case because the information yielded by tests is of knowable quality—and often of higher quality than other sources of in formation. According to the *Standards:* "The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route to broader and more equitable access to education and employment" (AERA/APA/NCME, 1999, p. 1). Because cut scores are the mechanism that results in category formation on tests, the importance of deriving defensible cut scores and their relevance to sound decision making are obvious. Again, according to the *Stan-dards:* "Verifying the appropriateness of the cut score or scores . . . is a critical element of the validity of test results" (p. 157).

Cross-Cutting Issues and General Considerations in Standard Setting

Several issues must be considered when setting performance standards regardless of which method is selected. Five such issues are described in the following paragraphs. A first consideration is the purpose of establishing standards in the first place. A common practice in all standard setting is to begin the session with an orientation for participants to the purpose of the task at hand. This orientation is a pivotal point in the process and provides the frame participants are expected to apply in the conduct of their work. Linn (1994) has suggested that standard setting can focus on one of four purposes: (1) exhortation, (2) exemplification, (3) accountability for educators, and (4) certification of student achievement. Depending on the purpose, the orientation to participants can differ substantially. For example, standard setting might involve exhortation. Using the policy rhetoric of higher standards, if the purpose were to "ratchet up expectations to world-class levels" for high school students in a state, the orientation provided to standard-setting participants might focus on describing the low level of challenge of previous content standards, the low bar set on previous state examinations, the evolving needs of the work force, and so on. An orientation like this, typically delivered by a person of relatively high status, would exhort participants to establish relatively high standards. By contrast, standard setting with an orientation of exemplification would focus more on providing concrete examples to educators of the competencies embedded in the content standards.

A second cross-cutting aspect of standard setting is the creation and use of *performance level labels* (PLLs). PLLs refer to the (usually) single-word terms used to identify performance categories; *Basic*, *Proficient*, and *Advanced* would be examples of such labels. Many such categorical labeling systems exist; a few examples are shown in Table 1. Though PLLs may have little technical underpinning, they clearly carry rhetorical value as related to the purpose of the standard setting. Such labels have the potential to convey a great deal in a succinct manner vis-à-vis the meaning of classifications that result from the application of cut scores. It is obvious from a measurement perspective that PLLs should be carefully chosen to relate to the purpose of the assessment, to the construct assessed, and to the intended, supportable inferences arising from the classifications.

A third issue—actually an extension of the concern with PLLs—is evident when *performance level descriptors* (PLDs) are created. PLD refers to the (usually) several sentences or paragraphs that provide fuller, more complete illustration of what performance within a particular category comprises. PLDs vary in their level of specificity, but have in common the verbal elaboration of the knowledge, skills, or attributes of test takers within a performance level. It is highly desirable for PLDs to be developed in advance of standard setting by a separate committee for approval by the appropriate policymaking body. Standard-setting participants then use these PLDs as a critical referent for their judgments. Sometimes, elaborations of the PLDs are developed by participants during a standard-setting procedure as a first step (i.e., prior to making any item or task judgments) toward operationalizing and internalizing the performance levels intended by the policy body. Sample PLDs, in this case those used for the NAEP Grade 4 reading assessment, are shown in Table 2.

There is an inherent tension in the creation of PLDs. Descriptions that provide too little specificity do not help illustrate or operationalize the performance categories. As such, they do not assist in communication to external audiences about the meaning of categorization at a given performance level. Descriptions that provide too much specificity by providing a detailed list of the knowledge and skills that a student at a given level possesses may be destined to pose validation problems. For example, suppose that very detailed descriptions are generated describing the specific knowledge and skills possessed by examinees in a category. Suppose further that actual categorical classifications will be based on examinees' total test scores. Under such a scenario, there will almost always be many instances in which a test taker demonstrates mastery of knowledge or skills *outside* the category to which he or she is assigned, and fails to demonstrate mastery of knowledge or skills for some elements *within* the performance category. This contradiction between the statement of knowledge and skills that examinees in a category are *supposed* to possess (as indicated in the PLDs) and the knowledge and skills that they *actually* possess (as indicated by observed test performance) makes validation of the PLDs problematic. Some researchers have attempted to solve this dilemma by crafting standard-setting procedures in which items are matched to performance level descriptions (see, e.g., Ferrara, Perie, & Johnson, 2002). Despite these efforts, the vexing issue of ensuring fidelity of PLDs with actual examinee performance is an area that remains one for which much additional work is needed.

Fourth, it has long been known that the participants in the standard-setting process are critical to the success of the endeavor and are a source of variability of standard-setting results. The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) provide guidance on representation, selection, and training of participants. For example, the *Standards* indicate that "a sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process were replicated" (p. 54). The *Standards* also recommend that "the qualifications of any judges involved in standard setting and the process by which they are selected" (p. 54) should be fully described and included as part of the documentation for the standard-setting process. The *Standards* also address training:

> Care must be taken to assure that judges understand what they are to do. The process must be such that well-qualified judges can apply their knowledge and experience to reach meaningful and relevant judgments that accurately reflect their understandings and intentions. (p. 54)

As with the development of PLDs, there is a tension present in the selection of standard-setting participants. While it is often recommended that participants have special expertise in the area for which standards will be set, in practice this can mean that standard-setting panels consist of participants whose perspectives are not representative of all practitioners in a field, all teachers at a grade level, and so on. Such a bias might be desirable if the purpose of standard setting is exhortation, though less so if the purpose of standard setting is to certify competence of students for awarding a high school diploma.

In addition, once standard-setting participants have been selected and trained and the procedure has begun, there is the matter of providing feedback to participants. Many standard-setting approaches comprise "rounds" or iterations of judgments. At each round, participants are provided various kinds of information to summarize their own variability, correspondence

Table 1. Sample Performance Level Labels

Labels	Source
Basic, Proficient, Advanced	National Assessment of Educational Progress
Starting Out, Progressing, Nearing Proficiency, Proficient, Advanced	*TerraNova*, 2nd ed. (CTB/McGraw-Hill)
Limited, Basic, Proficient, Accelerated, Advanced	State of Ohio Achievement Tests
Far Below Basic, Below Basic, Basic, Proficient, Advanced	State of California, California StandardsTests
Did Not Meet Standard, Met Standard, Commended Performance	State of Texas, Texas Assessment of Knowledge and Skills

Table 2. NAEP Performance Level Descriptors for Grade 4 Reading Tests

Performance Level Label	Performance Level Descriptor
Advanced	Fourth-grade students performing at the Advanced level should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. When reading text appropriate to fourth grade, they should be able to judge texts critically and, in general, give thorough answers that indicate careful thought. For example, when reading **literary** text, Advanced-level students should be able to make generalizations about the point of the story and extend its meaning by integrating personal experiences and other readings with ideas suggested by the text. They should be able to identify literary devices such as figurative language. When reading **informational** text, Advanced-level fourth-graders should be able to explain the author's intent by using supporting material from the text. They should be able to make critical judgments of the form and content of the text and explain their judgments clearly.
Proficient	Fourth-grade students performing at the Proficient level should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to fourth grade, they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connections between the text and what the student infers should be clear. For example, when reading **literary** text, Proficient-level fourth graders should be able to summarize the story, draw conclusions about the characters or plot, and recognize relationships such as cause and effect. When reading **informational** text, Proficient-level students should be able to summarize the information and identify the author's intent or purpose. They should be able to draw reasonable conclusions from the text, recognize relationships such as cause and effect or similarities and differences, and identify the meaning of the selection's key concepts.
Basic	Fourth-grade students performing at the Basic level should demonstrate an understanding of the overall meaning of what they read. When reading text appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences, and extend the ideas in the text by making simple inferences. For example, when reading **literary** text, they should be able to tell what the story is generally about—providing details to support their understanding—and be able to connect aspects of the stories to their own experiences. When reading **informational** text, Basic-level fourth graders should be able to tell what the selection is generally about or identify the purpose for reading it, provide details to support their understanding, and connect ideas from the text to their background knowledge and experiences.

with the group's ratings, or likely impact on the examinee population.

A complete treatment of selecting, training, and providing feedback to participants in standard setting is beyond the scope of this module. Readers are referred to the work of Raymond and Reid (2001) for further information on the selection, training, and evaluation of standard-setting participants, and to Reckase (2001) for more information on providing feedback to participants.

Finally, a fifth common issue is the necessity for standard-setting participants to form and rely on a conceptualization related to the examinee group to whom the standard(s) will apply. The need for such conceptualizations may have origins in the Nedelsky (1954) method in which standard setters are required to consider options that a hypothetical *F/D student* would recognize as incorrect. (According to Nedelsky, the F/D student is one who was on the borderline between passing and failing a course; hence, the notion of a point differentiating between a failing grade of "F" and a passing grade of "D.") Participants using an Angoff (1971) or derivative methodology form a conceptualization of the *minimally competent* examinee.

In contemporary standard setting, these often-hypothetical conceptualizations remain important, regardless of whether a particular method is considered to be "examinee centered" or "test centered" (Jaeger, 1989). For example, to use the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001), participants must consider at what point students in a certain performance category (e.g., *Basic*) or on the borderline between categories will have a specified probability of responding correctly. While standard-setting participants are often selected for their subject area expertise and knowledge of examinees to whom the test will be given, the abstract notion of an examinee within or between particular categories is still required for standard setting to proceed.

Standard-Setting Methods

According to the *Standards for Educational and Psychological Testing*, "There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility" (AERA/APA/NCME, 1999, p. 53). Recent advances in standard setting have added new approaches to the inventory of available methods. The new methods described in the following sections have some common advantages: they are generally more holistic (they require standard-setting participants to make holistic judgments about items or examinee test performance); they are intended to reduce the cognitive burden on participants; and they can be applied to a wide variety of item and task formats.

Before turning to a description of three such methods, we note two purposeful omissions in the following subsections. First, we do not review methods that would be highly appropriate for situations involving a mix of item formats and multiple cut scores (e.g., contrasting groups, borderline groups), but which have been described in previous modules (see Cizek, 1996a). Second, the following descriptions of each method generally focus on the procedures used to actually obtain one or more cut scores. Of course, much more is required of a defensible standard-setting process, including identification and training of appropriately qualified participants, effective facilitation, monitoring, and feedback to participants, and well conceived data collection to support whatever validity claims are made. A generic framework of steps required for standard setting has been put forth by Hambleton (1998) and is presented here as Table 3. However, each step warrants deeper attention in its own right, and readers interested in additional details on these topics are referred to other sources (e.g., Kane, 2001; Raymond & Reid, 2001; Reckase, 2001).

Bookmark Method

The Bookmark method is one of several item-mapping procedures developed in an attempt to simplify the cognitive task of standard setters who are required to consider performance-level descriptions, maintain appropriate conceptualizations of examinees within or between performance levels, and make probability estimates. First introduced by Lewis, Mitzel, and Green in 1996, the procedure has rapidly become widely used in K–12 education settings. Among the advantages of the Bookmark method are the comparative ease with which it can be applied by standard-setting participants, the fact that it can be applied to tests comprising both selected-response (SR, e.g., multiple-choice) and constructed-response (CR) items, and the fact that it can be used to set multiple cut scores on a single test.

The Ordered Item Booklet. The Bookmark procedure is so named because standard-setting participants identify cut scores by placing markers in a specially prepared test booklet. The distinguishing characteristic of the special test booklet is that it is prepared in advance with test items ordered by difficulty—easiest items first and hardest items last. This has come to be referred to as an *ordered item booklet* (OIB). The preparation of an OIB may seem simple enough in concept yet, until Lewis et al. (1996) introduced the idea, it had not been incorporated into a formal standard-setting method. The idea, however, instantly transformed standard setting into a classical psychophysics experiment in which a stimulus of gradually changing strength or form is presented to subjects who are given the task of noting the point at which a just-noticeable difference (JND) occurs. In the Bookmark procedure, participants begin with the knowledge that each succeeding item will be harder than (or at least as hard as) the one before; they are charged with noting one or more JNDs in the course of several test items in the OIB.

The ordering of MC items in an OIB is rather straightforward, particularly if a one-parameter logistic (1-PL) item response model (e.g., Rasch model) was used to obtain estimates of item difficulty. Whether a 1-PL, 2-PL, or 3-PL model is used, items are simply arranged in ascending *b*-value (i.e., item difficulty) order. When a test contains both SR and CR items, each CR item appears several times in the booklet—once for each of its score points. For a given CR item, the item prompt, the rubric, and sample examinee responses illustrating the score point(s) are also provided to standard setters. The OIB is formatted with only one item (or CR score point) per page.

The OIB can be composed of any collection of items that is representative of

Table 3. Generic Steps in Setting Performance Standards

Step	Description
1	Select a large and representative panel.
2	Choose a standard-setting method; prepare training materials and standard-setting meeting agenda.
3	Prepare descriptions of the performance categories (i.e., PLDs).
4	Train participants to use the standard-setting method.
5	Compile item ratings or other judgments from participants and produce descriptive/summary information or other feedback for participants.
6	Facilitate discussion among participants of initial descriptive/summary information.
7	Provide an opportunity for participants to generate another round of ratings; compile information and facilitate discussion as in Steps 5 and 6.
8	Provide a final opportunity for participants to review information and arrive at final recommended performance standards.
9	Conduct an evaluation of the standard-setting process, including gathering participants' confidence in the process and resulting performance standard(s).
10	Assemble documentation of the standard-setting process and other evidence, as appropriate, bearing on the validity of resulting performance standards.

Source: Adapted from Hambleton (1998).

the range of content, item types, and summary statistical characteristics of a typical test form. An OIB need not consist only of items that appear in an actual test; it can have more or fewer items than an operational test booklet. However, it is important that the OIB fully represent the breadth and depth of content to which examinees will be exposed in order for standard-setting participants to understand more clearly the precise ability level needed to achieve a particular standard. Thus, it is most common for the OIB to comprise an intact test form. One advantage of using an operational form is that participants evaluate the test on which the standard will be set, as opposed to reviewing some items to which examinees may never actually be exposed.

An example of a page from an OIB is shown in Figure 1. The example is taken from a high-stakes reading test administered to high school students in a large midwestern state. Detailed PLDs, based on the state's content standards, were developed in advance and used by standard-setting participants (n = 20) to identify three cut scores separating four performance levels: *Advanced, Proficient, Basic,* and *Below Basic.*[3]

The boldfaced number in the upper-right corner of the page is simply pagination; the item in this example appeared on page 35 of the OIB. The next information provided is the item's position in the intact test form (it was item number 22) and the item response theory (IRT) ability level required to have a .67 probability of answering the item correctly—in this case 1.725. Information preceding the item indicates that it is one of a set of items associated with a passage titled "Yellowstone." (A collection of all passages used in the test would be supplied to participants as a separate booklet for their use during standard setting.) An asterisk by option C indicates the correct response. Had this been a CR item, the prompt would have been followed by a sample response at a particular score point; in the full OIB, the prompt and an associated sample response would appear once for each of its non-zero score points, distributed throughout the OIB in order of the difficulty of obtaining the particular score point (or higher).

Probability Judgments in the Bookmark Approach. In using the Bookmark method, participants must make a probability judgment. In essence, they must concern themselves with a question such as, "Is it likely that an examinee on the borderline between categories X and Y will answer this MC item correctly (or earn this CR item point)?" Obviously, to actually implement the Bookmark method the task becomes one of defining "likely." In practice, most applications of the Bookmark method employ a 67% likelihood of the correct response (for SR items), or of obtaining at least a particular score point (for CR items). Standard-setting participants are instructed to place a marker in their OIB on the page (i.e., item) *immediately after* the page at which, in their opinion, the likelihood criterion applies, that is, to place their bookmarks at the first point in the booklet at which they believe examinees' probability of making the desired response drops below .67. It is important to note that this point is *not* the cut score in the sense that the point at which the marker is placed cannot be translated into a raw cut score by counting the number of items preceding it. Rather, as will be shown in the next section, the cut score will be determined by obtaining the scale value (often an IRT ability estimate) corresponding to a .67 probability of answering the item correctly.

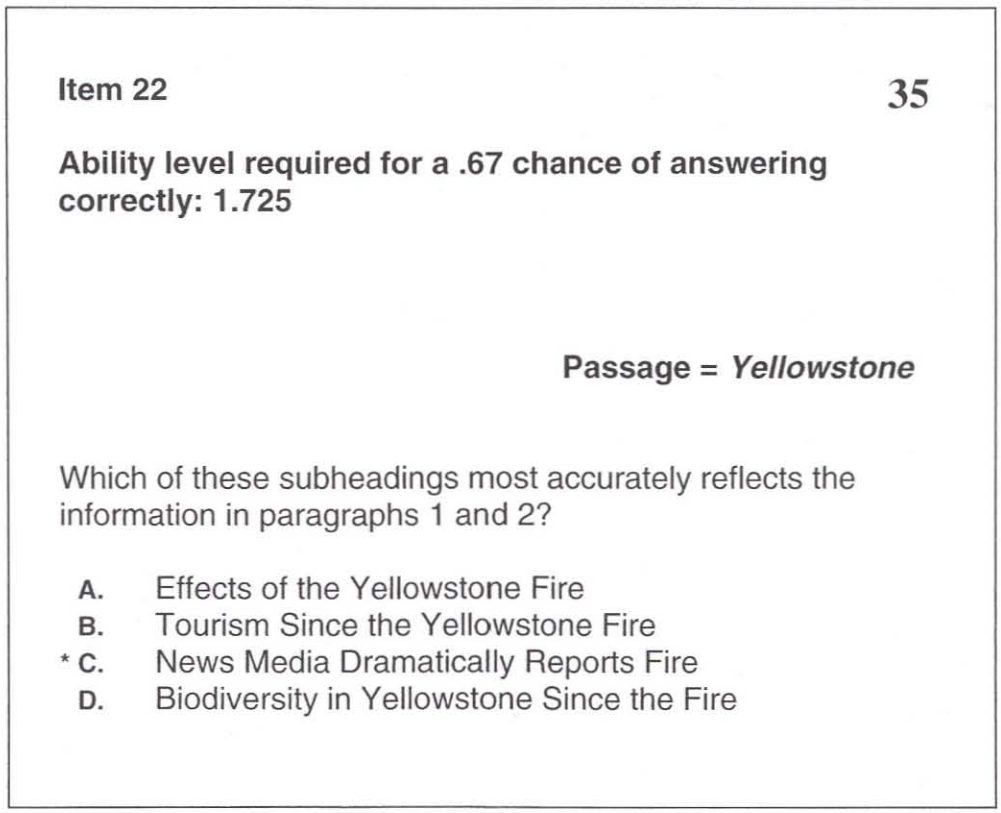

FIGURE 1. Sample page from ordered item booklet.

The particular likelihood used—in this case .67—is referred to as the *response probability* (RP). According to Mitzel et al. (2001), an RP of .67 can be interpreted in the following way: "For a given cut score, a student with a test score at that point will have a .67 probability of answering an item also at that cut score correctly" (p. 260). However, the use of other RPs has been investigated. Huynh (2000) suggested that the RP which maximized the information function of the test would produce the optimum decision rule. For a two-parameter IRT model, Huynh found that an RP of .67 maximized this function.

Wang (2003) concluded that an RP of .50 is preferable when the Rasch (i.e., 1-PL) scaling model is used. The choice of .50 in the Rasch model context has certain mathematical advantages over .67 in that the likelihood of a correct response is exactly .50 when the examinee ability is equal to the item difficulty.

Issues related to selection of the most appropriate RP remain, however. Whether standard-setting participants can use any particular RP value more effectively than another and whether they can understand and apply the concept of RP more consistently and accurately than they can generate probability estimates using, for example, a modified-Angoff approach remain topics for future research.

Psychometric Foundations of the Bookmark Approach. As originally described, the Bookmark method employs a three-parameter logistic (3-PL) model for SR items and a two-parameter partial-credit (2PPC) model for CR items. However, an alternative approach using a 1-PL (i.e., Rasch) model for both SR and CR items is also frequently used in practice. Both approaches are described in this section beginning with a brief explication of the Bookmark method as originally proposed.

As indicated previously, standard-setting participants express their judgments by placing a marker in the OIB on the page after the last item that they believe an examinee who is just barely qualified for a particular classification (e.g., *Proficient*) has a .67 probability of answering correctly. These judgments are translated into cut scores by noting the examinee ability associated with a .67 probability of a correct response and then translating that ability into a raw score. As originally described by Mitzel et al. (2001), the probability of a correct response (P_j) for an SR item is a function of examinee ability (θ), item difficulty (b_j), item discrimination (a_j), and a threshold or chance variable (c_j) in accordance with the fundamental equation of the 3-PL model:

$$P_j(\theta) = c_j + (1 - c_j)/\{1 + \exp[-1.7a_j(\theta - b_j)]\}, \quad (1)$$

where exp represents the natural logarithm e (2.71828 ...) raised to the power of the expression to the right. However, Mitzel et al. (2001) set the threshold or chance parameter (c_j) equal to zero, reducing Equation 1 to

$$P_j(\theta) = 1/\{1 + \exp[-1.7a_j(\theta - b_j)]\}, \quad (2)$$

or essentially a 2-PL model.

For dichotomously scored (i.e., SR) items, the basic standard-setting question is whether or not an examinee just barely categorized into a given performance level would have a .67 chance of answering a given SR item correctly. Thus, starting with a probability of .67 and solving Equation 2 for the ability (θ) needed to answer an item correctly, we obtain the following:

$$\theta = b_j + .708/1.7a_j. \quad (3)$$

For CR items, the situation becomes somewhat more complicated. Mitzel et al. (2001) used the two-parameter generalized partial-credit model (Muraki, 1992). This model, shown in Equation 4, presents the probability of obtaining a given score point (c), given some ability level (θ), as a function of the difficulty of the various score points (b_{i0} to b_{ik}) and the item discrimination (a_i):

$$P_{ik}(\theta) = \frac{\exp\left[\sum_{v=0}^{k} a_i(\theta - b_{iv})\right]}{\sum_{c=0}^{m_i} \exp\left[\sum_{v=0}^{c} a_i(\theta - b_{iv})\right]}. \quad (4)$$

Mitzel et al. (2001) note that the Bookmark procedure can also be implemented under other IRT models, such as the 1-PL (Rasch) model. This particular application of the Bookmark procedure begins with a basic expression of the Rasch model for dichotomous items (cf., Wright & Stone, 1979; Equation 1.4.1):

$$P(X = 1|\theta_v, \delta_i) = \exp(\theta_v - \delta_i)/[1 + \exp(\theta_v - \delta_i)], \quad (5)$$

where

θ_v = ability (theta estimate) of an examinee;

δ_i = difficulty of item i; and

exp = natural logarithm raised to the power inside the parentheses.[4]

Allowing the expression on the right of Equation 5 to equal .67 and solving for θ_v, we obtain the following:

$$\theta_v = \delta_i + .708, \quad (6)$$

which is very similar to Equation 3 except for the omission of the a parameter, which is the distinguishing characteristic of the 2-PL model. Thus, the Rasch ability level required for an examinee to have a .67 probability of answering a given SR item correctly would be .708 logits *greater than* the difficulty of the item.

When a test comprises CR items, the derivation of the ability level necessary to obtain a given score point is somewhat more complex than for SR items. Indeed, it is necessary to calculate a system of probabilities for each CR item (i.e., a probability for each score point). To accomplish this, a partial-credit model is commonly used. According to this model, the likelihood (π_{nix}) of a person (n) with a given ability (θ_n) obtaining a given score (x) on an item (i) with a specified number of steps (j) is shown in Equation 7 (taken from Wright & Masters, 1982, Equation 3.1.6):

$$\pi_{nix} = \frac{\exp\sum_{j=0}^{x}(\theta_n - \delta_{ij})}{\sum_{j=0}^{m_i}\exp\sum_{j=0}^{j}(\theta_n - \delta_{ij})}, \quad (7)$$

where x is the value of the score point (0, 1, 2, 3, etc.) in question, and m_i is the final step. The numerator in Equation 7 refers only to the steps completed for the score point x, while the denominator includes the sum of all $m_i + 1$ possible numerators.

Determining a Cut Score Using the Bookmark Method. The following example illustrates the application of the Bookmark procedure. The items shown in Table 4 are drawn from a report by Schagen and Bradshaw (2003) regarding a national reading test given to 11-year-olds in Great Britain. The test consisted of 27 SR items and 10 CR items. Of the 10 CR items, seven were worth 2 points each, and three were worth 3 points each, for a total of 50 points for the entire test. Twelve participants evaluated the OIB represented in Table 4 and rendered their bookmark placements for a *minimal* student (Level 3). Those judgments are shown in Table 5.

Table 4. Ordered Booklet Item Parameters and Associated Theta Values

Page	Item	Difficulty (b)	Discrim. (a)	Theta @ RP = .67	Page	Item	Difficulty (b)	Discrim. (a)	Theta @ RP = .67
1	19	−3.395	0.493	−2.550	26	32	−0.341	0.869	0.138
2	13	−2.770	0.997	−2.352	27	29.1	−0.333	0.667	0.160
3	1	−2.757	1.441	−2.468	28	11	−0.133	0.494	0.710
4	22	−2.409	0.461	−1.505	29	37.1	−0.120	0.515	0.120
5	4	−2.282	0.527	−1.492	30	10	−0.063	0.402	0.973
6	2	−2.203	0.607	−1.517	31	31.2	−0.052	0.817	0.940
7	12	−2.141	0.503	−1.313	32	16	0.107	0.316	1.425
8	3	−1.781	0.520	−0.980	33	6	0.247	0.866	0.728
9	14	−1.737	0.931	−1.290	34	36	0.312	0.421	1.301
10	31.1	−1.710	0.817	−1.240	35	24	0.396	0.489	1.248
11	23	−1.454	0.778	−0.919	36	35.1	0.469	0.586	1.060
12	21	−1.444	0.845	−0.951	37	26.2	0.558	0.563	1.280
13	7	−1.122	0.953	−0.685	38	30.2	0.806	0.600	2.220
14	20.1	−1.044	0.743	−0.830	39	17	0.931	0.724	1.506
15	28	−0.973	0.770	−0.432	40	37.2	1.099	0.515	1.920
16	30.1	−0.942	0.600	−0.420	41	18	1.390	0.572	2.118
17	34.1	−0.935	0.657	−0.270	42	29.2	1.513	0.667	2.190
18	15	−0.873	0.567	−0.138	43	26.3	1.519	0.563	3.180
19	9	−0.833	0.863	−0.350	44	34.2	1.541	0.657	2.750
20	8	−0.724	0.901	−0.262	45	27.1	2.062	0.292	2.450
21	25.1	−0.703	0.750	0.010	46	25.2	2.293	0.750	3.310
22	5	−0.500	0.595	0.200	47	37.3	2.384	0.515	4.160
23	26.1	−0.424	0.563	−0.270	48	35.2	2.479	0.586	3.900
24	20.2	−0.422	0.743	0.840	49	29.3	3.149	0.667	4.420
25	33	−0.379	0.828	0.124	50	27.2	3.174	0.292	6.440

Note: CR items have multiple entries. For example, Item 37 has three score points, shown as score point 37.1 (OIB page 29), 37.2 (OIB page 40) and 37.3 (OIB page 47).
Source: Adapted from Schagen and Bradshaw (2003).

The cut score is based on the mean theta at the associated response probability (theta @ RP = .67). In this instance, the mean theta value of −1.594 corresponds to a raw score of 15.25. Because fractional raw scores are not possible, the operational cut score would need to be rounded to a possible score point, such as 15 or 16, depending on the rounding rules in place, though it should be noted that a student who had earned a raw score of 15 would have an ability less than the target value of −1.594.

It should also be noted that participants selected items on the second, fifth, and sixth pages of the OIB (Items 13, 4, and 2, respectively). If none of the participants went farther than page 6 in the booklet, it might seem reasonable that the cut score for the minimal level should be no more than 6 points. However, the Bookmark procedure focuses on the student ability level associated with the 67% likelihood of answering Item 2, 4, or 13 (the ones identified by the participants as marking the boundary between minimal and the next lower level). It is on those ability levels, not the page numbers or cumulative number of items, that the cut score is set. The student who has a 67% likelihood of answering Item 2 correctly also has a slight chance of answering subsequent items correctly or obtaining scores of 2 or 3 on moderately difficult CR items. The expected score for the student at the just barely minimal level is the aggregate of expected scores on all 37 items in the test. For this particular test, based on the average of these participants' estimates, that expected raw score is somewhere between 15 and 16.

To summarize this application of the Bookmark method, 12 standard-setting participants made judgments about the location of the *minimal* achievement level by placing bookmarks in their OIBs. These judgments are shown in the column labeled "Item Number" in Table 5. The relationships for each item between page number and ability required to reach that level (with a 67% likelihood) are shown in Table 4. The page numbers supplied by the participants were translated into ability estimates using the data in Table 4. These ability estimates were then averaged to determine the mean ability estimate of a student just barely at the *minimal* level. That ability level was then converted to a raw score using standard, commercially available 3PL model software.

Angoff Variations

Originally proposed by Angoff (1971) and described elsewhere (see Cizek, 1996a), the Angoff approach has produced many variations which have adapted this most thoroughly researched and still widely used method to evolving assessment contexts and challenges. Just as the previously described Bookmark approach was developed in an attempt to reduce the complexity of the cognitive task facing standard-setting participants, so too does a derivative of the Angoff procedure referred to as the *Yes/No method* by Impara and Plake (1997). The essential question that must be addressed by standard-setting participants can be answered "Yes" or "No." According to Impara and Plake, participants are directed to

Table 5. Summary of Participants' Bookmark Placements for Level 3 (Minimal)

Participant	Item Number	Page Number in OIB	Theta @ RP = .67
A	2	6	−1.517
B	4	5	−1.492
C	4	5	−1.492
D	2	6	−1.517
E	2	6	−1.517
F	2	6	−1.517
G	13	2	−2.352
H	4	5	−1.492
I	2	6	−1.517
J	13	2	−2.352
K	2	6	−1.517
L	2	6	−1.517
Mean			−1.594

Source: Adapted from Schagen and Bradshaw (2003).

> read each item [in the test] and make a judgment about whether the borderline student you have in mind will be able to answer each question correctly. If you think so, then under Rating 1 on the sheet you have in front of you, write in a Y. If you think the student will not be able to answer correctly, then write in an N. (pp. 364–365)

In essence then, the Yes/No method is highly similar to the first Angoff (1971) approach. In his oft-cited chapter on scaling, norming, and equating, Angoff described two variations of a standard-setting method. While his second suggestion came to be known as the widely used Angoff method, Angoff first suggested that standard setters simply judge whether or not a hypothetical minimally acceptable person would answer an item correctly. According to Angoff,

> 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." (pp. 514–515)

Implementing the Yes/No Method. The basic procedures for implementing the Yes/No method follow those for most common standard-setting approaches. To begin, qualified participants are selected and are oriented to the standard-setting task. They are often grounded in the content standards upon which the test was built; they may be required to take the test themselves, and they discuss the relevant competencies and characteristics of the target population of examinees for whom the performance levels are to be set. After discussion of the borderline examinees, participants are asked to make performance estimates for a group of examinees in an iterative process over two or more "rounds" or ratings.

Typically, in a first round of performance estimation, participants using the Yes/No method rate a set of operational items often comprising an intact test form. At the end of Round 1, each participant would be provided with feedback on their ratings in the form of information about how their ratings compared to actual examinee performance or to other participants' ratings. A second round of yes/no judgments on each item follows as participants re-review each item in the test. If not provided to them previously, at the end of the second round of judgments, participants would receive additional information regarding how many examinees would be predicted to pass/fail based on their participants' judgments (i.e., impact data). Regardless of how many rounds of ratings occur, calculation of the final recommended passing score would be based on data obtained in the final round.

Extended Angoff Method. Although an extension of the Yes/No method to contexts with polytomously scored items or a mix of SR and CR formats has not been attempted, another variation of Angoff's (1971) basic approach has been created to address tests that include CR items. Hambleton and Plake (1995) describe what they have labeled an *extended Angoff procedure.* In addition to providing traditional probability estimates of borderline examinee performance for each SR item, participants also estimate the number of scale points that they believe borderline examinees will obtain on each CR task in the assessment. Cut scores for the extended Angoff approach are calculated in the same way as with traditional Angoff methods, although, as Hambleton (1998) notes, more complex weighting schemes can also be used for combining components in a mixed-format assessment.

Calculation of Yes/No and Extended Angoff Cut Scores. Table 6 presents hypothetical data for the ratings of 20 items by six participants in two rounds of ratings using the Yes/No and extended Angoff methods. The table has been prepared to illustrate calculation of cut scores that would result from use of the Yes/No method alone for a set of dichotomously scored SR items (i.e., the first 12 items listed in the table), the extended Angoff method alone for a set of polytomously scored CR items (the last eight items in the table), or a combination of Yes/No and extended Angoff (for the full 20-item set). For this set of items the CR items were scored on a 1–4 scale.

The means for each participant and each item are also presented for each round. Using the Round 2 ratings shown in Table 6, the recommended Yes/No passing score for the SR item test would be approximately 58% of the total raw score points (.58 × 12 items), or approximately 7 out of 12 points possible. The recommended passing score on the CR item test would be 21 out of a total of 32 possible score points (2.69 × 8 items). A recommended passing score for the 20-item test comprising a mix of SR and CR items would be approximately 28 of the 44 total possible raw score points [(.58 × 12) + (2.69 × 8)]. (See Hambleton & Plake, 1995 and Talente, Haist, & Wilson, 2003 for additional information on setting standards for complex performance assessments.)

Research on the Yes/No Method. One of the appealing features of the Yes/No

Table 6. Hypothetical Data and Examples of Yes/No and Extended Angoff Standard-Setting Methods

	Participant ID Number						
Item No.	1	2	3	4	5	6	Means
1	1	0	0	1	0	1	0.50
	1	1	0	0	0	1	0.50
2	0	0	0	0	0	0	0.00
	0	0	0	1	0	0	0.17
3	1	1	0	1	1	1	0.83
	1	1	0	1	1	1	0.83
4	1	1	1	1	1	1	1.00
	1	1	1	1	1	1	1.00
5	0	0	0	0	0	0	0.00
	0	0	0	0	0	0	0.00
6	0	0	0	0	0	0	0.00
	0	0	0	0	0	0	0.00
7	1	1	1	1	1	1	1.00
	1	1	1	1	1	1	1.00
8	1	1	1	1	1	1	1.00
	1	1	1	1	1	1	1.00
9	1	1	1	1	1	1	1.00
	1	1	1	1	1	1	1.00
10	1	1	1	0	1	1	0.83
	1	1	1	0	1	1	0.83
11	0	0	0	0	0	0	0.00
	0	0	0	0	0	0	0.00
12	0	0	1	0	0	0	0.17
	1	0	1	1	1	0	0.67
Means	.58	.50	.50	.50	.50	.58	.53
	.67	.58	.50	.58	.58	.58	**.58**
13	2	3	2	2	3	1	2.17
	3	3	3	3	3	2	2.83
14	1	2	1	2	2	1	1.50
	2	2	2	2	3	2	2.17
15	2	2	2	2	2	2	2.00
	3	3	3	3	3	2	2.83
16	3	3	2	2	3	2	2.50
	3	3	3	3	3	3	3.00
17	1	1	2	1	2	1	1.33
	2	2	2	2	2	1	1.83
18	2	3	3	2	3	2	2.50
	3	3	3	3	3	2	2.83
19	3	2	2	2	3	2	2.33
	3	3	3	3	3	3	3.00
20	2	3	3	2	3	2	2.50
	3	3	3	3	3	3	3.00
Means	2.00	2.38	2.13	1.88	2.63	1.63	2.10
	2.75	2.75	2.75	2.75	2.88	2.25	**2.69**

Note: The upper and lower entries in each cell represent participants' first and second round ratings, respectively; values in bold are Round 2 means for SR and CR items, respectively.

method is its simplicity. In typical implementations of modified Angoff procedures, participants must maintain a concept of a group of hypothetical examinees and must estimate the proportion of that group which will answer an item correctly. Clearly, this is an important—though difficult—task. Impara and Plake (1998) found that the Yes/No method ameliorated some of the difficulty of the task. They reported that:

> We believe that the yes/no method shows substantial promise. Not only do panelists find this method clearer and easier to use than the more traditional Angoff probablility estimation procedures, its results show less sensitivity to performance data and lower within-panelist variability. Further, panelists report that the conceptualization of a typical borderline examinee is easier for them than the task of imagining a group of hypothetical target candidates. Therefore, the performance standard derived from the yes/no method may be more valid than that derived from the traditional Angoff method. (p. 336)

As Impara and Plake (1998) have demonstrated, even teachers who were familiar with an assessment and with the examinees taking the assessment were not highly accurate when asked to predict the proportion of a group of borderline students who would answer an item correctly. The Yes/No method simplifies the judgment task by reducing the probability estimation required to a dichotomous outcome.[5]

There are two alternative ways in which the Yes/No method can be applied. One variation requires participants to form the traditional conceptualization of a hypothetical borderline examinee; the other requires participants to reference their judgments with respect to an actual examinee on the borderline between classifications (e.g., between Basic and Proficient). In a comparative trial of the Yes/No method with a modified Angoff approach, Impara and Plake (1997) asked participants using the Yes/No method to think of one *actual* borderline examinee with whom the participant was familiar instead of conceptualizing a group of *hypothetical* examinees. Keeping this *actual* person in mind, participants were then asked to determine whether the examinee would answer each item correctly. The results showed that although the final standard was similar for participants using the Angoff method and the Yes/No method, the variance of the ratings with the Yes/No method was smaller and the participants' scores were more stable from Round 1 to Round 2. Participants reported that thinking of an *actual* examinee when rating the items was easier than thinking of a group of *hypothetical* examinees.

The relative cognitive simplicity of the Yes/No method identified by Impara and Plake was also reported by Chinn and Hertz (2002). They report that participants found the yes/no decisions easy to make because "they were forced to decide between a yes or a no rather than estimate performance from a range of estimates," whereas participants using a modified Angoff method "commented that determining the proportion of candidates who would answer each item correctly was difficult and subjective" (p. 7). However, in contrast to the attractive stability of the participants' ratings observed by Impara and Plake (1998), Chinn and Hertz found that there was greater variance in ratings using the Yes/No method. They hypothesize that this may be due to design limitations and several departures from the methodology used by Impara and Plake including their selection of participants, instructions, and level of discussion about the process.

To date the Yes/No method has only been applied in contexts where the outcome is dichotomous (i.e., with multiple-choice or other SR-format items which will be scored as correct or incorrect).

Holistic Methods

Increasingly, large-scale assessments have incorporated a mix of item formats in order to tap more fully the constructs that are measured by those tests and to avoid one common validity threat known as construct underrepresentation. While tests comprising SR-format items exclusively may have been more common in the past, newer tests often comprise short-response items, essays, show-your-work, written reflections, grid-in response format, and other test construction features for which standard-setting methods designed for SR tests are not amenable.

Assessment specialists have responded by proposing a variety of methods for setting performance standards on tests comprising exclusively CR items (e.g., a writing test) or a mix of SR and CR formats (e.g., a mathematics test). Several of these methods can be termed "holistic," in that they require participants to focus judgment on a sample or collection of examinee work greater than a single item or task at a time. Though a number of methods satisfy this characteristic, we are aware, too, that differences between these methods can defy common classification. With that caveat, we note several examples of more holistic methods, then we provide greater detail on a single implementation of one such procedure.

Examples of Some Holistic Methods. One such method that would be considered more holistic has been proposed by Plake and Hambleton (2001) (although the developers described their method as "analytic judgment"). The method was developed for tests that include polytomously scored performance tasks and other formats, resulting in a total test comprising different components. To implement the method, panelists review a carefully selected set of materials for each component, representing the range of actual examinee performance on each of the questions comprising the assessment (although examinees' scores are not revealed to the panelists). Panelists then classify the work samples according to whatever performance levels are required (e.g., *Basic, Proficient,* and *Advanced*). Plake and Hambleton used even narrower categories within these performance levels, which they called low, middle, and high (e.g., low-*Basic,* middle-*Basic,* high-*Basic*). Although Plake and Hambleton suggested alternative methods for calculating the eventual cut scores, a simple averaging approach appeared to work as well as the others. The averaging approach consisted of taking all papers classified by participants into what were called borderline categories. For example, the cut score distinguishing *Basic* from *Proficient* was obtained by averaging the scores of papers classified into the high-*Basic* and low-*Proficient* borderline categories.

Loomis and Bourque (2001) have described a similar approach to that of Plake and Hambleton (2001) in what they call a *paper selection method.* They also describe another similar approach, which they term the *booklet classifica-*

tion method; the latter method differs in essence from the component-based methods in that it requires participants to engage in the sorting/ classification task at the level of an entire test booklet. What can be termed "holistic" methods have also been proposed by Jaeger (1995) in a *judgmental policy capturing* approach and by Putnam, Pence, and Jaeger (1995) in the *dominant profile* method. For additional information on any of these methods, readers should consult the corresponding original sources listed. In the following paragraphs, we provide detail on one holistic method as an example of the characteristics of such an approach.

The Body of Work Method. One fairly well known holistic method is the *Body of Work (BoW)* method, proposed by Kingston, Kahl, Sweeney, and Bay (2001). The BoW method differs somewhat from other holistic methods in its calculation of cut scores. Rather than taking simple means of borderline groups (which may be skewed if not moderated to account for different numbers of examinees in the two groups), Kingston et al. (2001) employ a logistic regression to derive cut scores. As with many standard-setting methods, a number of variations of Kingston et al.'s basic suggestion have been implemented, and comprise what we refer to generally as a *holistic work sample* method. Information in the following paragraphs is relevant for obtaining cut scores using this genre of standard-setting methods, regardless of the label applied.

Holistic approaches typically present large numbers of intact student work samples to participants. Typically, these student work samples have been scored prior to standard setting, but the individual scores are not provided to participants during the judgment process. Instead, participants rate each work sample holistically and classify it into one of the required categories (e.g., *Below Basic, Basic, Proficient,* or *Advanced*). In preparation for the standard-setting meeting, as many as 1,000 scored student work samples may be reviewed by standard-setting facilitators; from that number, 40 to 50 samples to represent the range of total scores may be selected.

Consider, for example, a language arts test consisting of two essays, a revise-and-edit task, and two reading passages with both SR and CR items, with a total score of 50 points. Selecting 40 student work samples would entail some decisions about which score points to leave in and which to leave out, since 40 work samples can represent, at most, 40 different score points. These 40 or so work samples are presented to participants who sort them into the categories such as the four performance levels named previously. Participants may then discuss their decisions in small groups or in a large group, and may modify some of their decisions before submitting their judgments to the facilitators. Where these within-round discussions occur, the rounds are sometimes subdivided into Round 1.1, 1.2, 2.1, 2.2, and so on. Following Round 1 and data analysis, preliminary cut scores are identified. In this case, there would be three cut scores, one to separate *Below Basic* from *Basic*, one to separate *Basic* from *Proficient*, and one to separate *Proficient* from *Advanced*. At this point, it will become evident that some score points are beyond consideration as possible cut scores.

In the current example, if no participant identified a work sample with a total score below 17 as belonging to the *Basic* category, then work samples with scores of 16 and below would be eliminated from further consideration, and additional work samples would be brought into the mix in Round 2 to augment the likely regions of the cut scores. For this reason, Round 1 is sometimes referred to as *range finding,* and Round 2 is referred to as *pinpointing* (see Kingston et al., 2001, pp. 226–230).

In Round 2, participants may reexamine some of the Round 1 work samples plus additional work samples that fill in any gaps in the ranges of the preliminary cut scores, or they may review all new work samples, selected on the basis of Round 1 results. Similarly, by Round 3, the range of scores represented in student work samples may be further curtailed.

To illustrate a holistic standard-setting approach, consider the 50-point language arts test described above. Twenty participants have rated 40 student work samples with scores ranging from 13 to 50. Participants do not know the scores of any of the work samples. The facilitators have purposely eliminated work samples with scores below 13, based on preliminary research. During Round 1, the 20 participants entered a total of 360 ratings, an average of 18 ratings per participant, though the rate varies considerably. Similarly, some work samples have been rated more times than others. Figure 2 shows the results of Round 1.

Each category (*Below Basic, Basic, Proficient, Advanced*) is represented by a score distribution. These distributions overlap to a considerable degree. Indeed, not only do some ratings for *Basic* overlap *Proficient* but also *Advanced.* This degree of overlap is not uncommon in Round 1 of a holistic rating procedure, and it frequently occurs in later rounds in holistic rating with certain kinds of assessments (e.g., those for alternate assessments for students with special needs).

There are three vertical lines in Figure 2, each representing a likely cut score: C1, C2, and C3. C1, for example, is placed where the *Below Basic* distribution crosses the *Basic* distribution. In a BoW application, this point would also correspond to the value yielded by logistic regression, which searches for the point at which the likelihood of being classified as *Basic* reaches 50%. This is at about 20 raw score points. Below 20 points, the work sample is more likely to be classified as *Below Basic*. At or above 20 points, the work sample is more likely to be classified as *Basic.* A similar shift occurs at about 29 points (*Basic* to *Proficient*) and again at about 39 points (*Proficient* to *Advanced*).

The computed cut scores will depend on the analytical method that accompanies the particular holistic method used. As noted above, the BoW method uses logistic regression to determine the point at which the likelihood of a particular classification reaches or first exceeds 50%. The analytic judgment method (Plake & Hambleton, 2001) would have subdivided the groups into high-*Basic,* low-*Proficient,* and so on, determined the mean scores for each of these borderline groups, and then produced a cut score equal to the midpoint between two adjacent borderline group means. Similarly, one might simply calculate the mean (or median) for each category and then calculate the midpoint between two adjacent category means to derive a cut score.

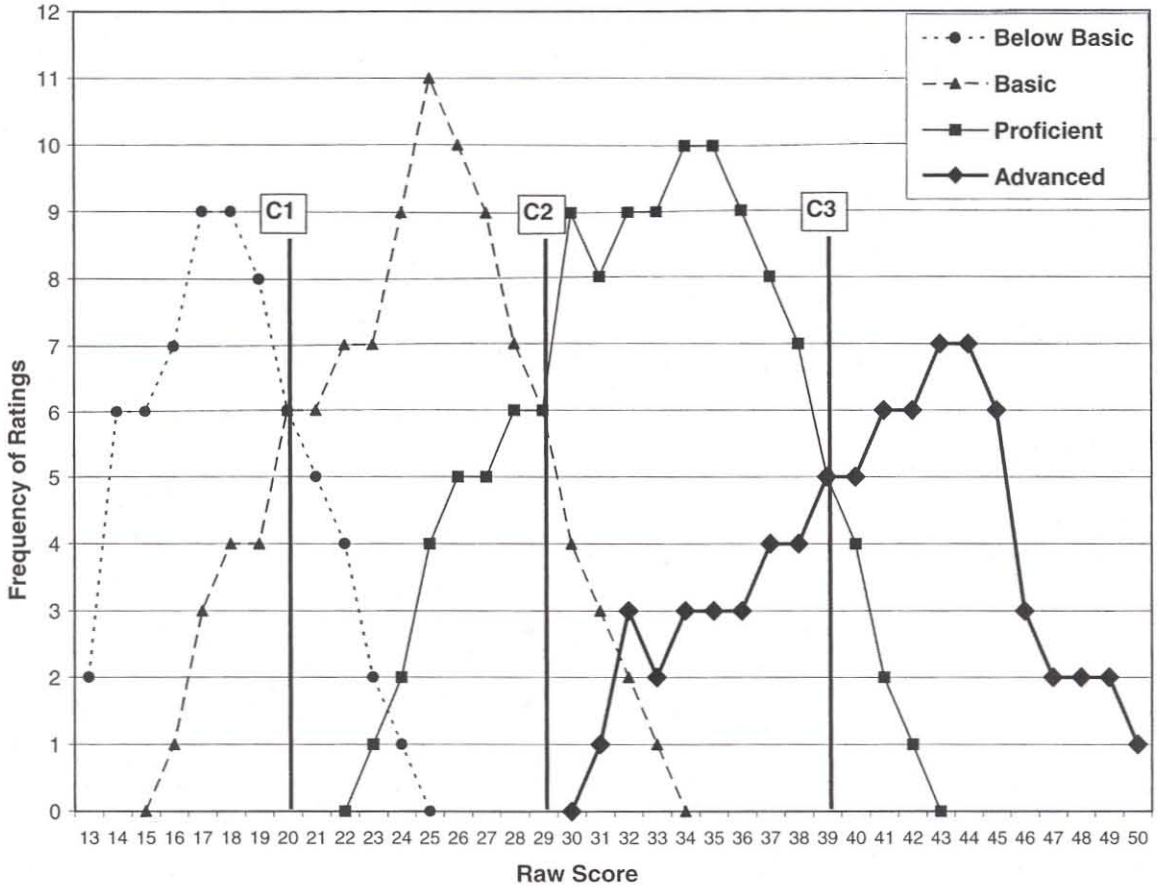

FIGURE 2. Results of holistic rating Round 1.

When using holistic approaches, decisions about whether and when to share student work sample scores, overall distributions of scores (i.e., impact data), item difficulty, and other data are made prior to the standard-setting activity. Typically, item and score data are shared after Round 1, and impact data are shared after Round 2. However, in some cases, impact data are also shared after Round 1.

Evaluating the Standard-Setting Process

Although not strictly a method itself, it is important that any standard-setting process gather evidence bearing on the manner in which any particular approach was implemented and the extent to which participants in the process were able to understand, apply, and have confidence in the eventual performance standards (Cizek, 1996b). Thus, evaluation of the standard-setting process can be thought of as an aspect of each method described previously in this module. Equal attention must be devoted to planning the standard-setting evaluation, *a priori*, as is given to carrying out the standard-setting procedure itself.

The evaluation of standard setting is a multifaceted endeavor. It can be thought of as beginning with a critical appraisal of the degree of alignment between the standard-setting method selected and the purpose and design of the test, the goals of the standard-setting agency, and the characteristics of the standard setters. This match should be evaluated by an independent body (such as a technical advisory committee) acting on behalf of the standard-setting agency. Evaluation continues with a close examination of the application of the standard-setting procedure: To what extent did it adhere faithfully to the published principles of the procedure? Did it deviate in unexpected, undocumented ways? If there are deviations, are they reasonable adaptations, specified and approved in advance, and consistent with the overall goals of the activity? A measure of the degree to which individual standard-setting participants converge from one round to the next is yet another part of the evaluation.

These aforementioned evaluations are external in nature. However, on-site evaluations of the process of standard setting, by the participants themselves, serve as an important internal check on the validity and success of the process. Typically, two evaluations are conducted during the course of a standard-setting meeting. A first evaluation normally occurs after initial orientation of participants to the process, training in the method, and (when appropriate) administration to participants of an actual test form. This first evaluation serves as a check on the extent to which participants have been adequately trained, understand key conceptualizations and the task before them, and have confidence that they will be able to apply the selected method. A second evaluation is ordinarily conducted at the conclusion of the standard-setting meeting. Commonly, both evaluations consist of a series of survey questions. A sample end-of-meeting survey is shown in Figure 3.

It should be noted that the format of the items in the survey shown in Figure 3 requires only an "Agree" or "Disagree" check mark from respondents. Because

Directions: Please check "Agree" or "Disagree" for each of the following statements and add any additional feedback on the process at the bottom of this page.

	Statement	Agree	Disagree
1	The orientation provided me with a clear understanding of the purpose of the meeting.		
2	The workshop leaders clearly explained the task.		
3	The training and practice exercises helped me understand how to perform the task.		
4	Taking the test helped me to understand the assessment.		
5	The performance level descriptions were clear and useful.		
6	The large and small group discussions aided my understanding of the process.		
7	The time provided for discussions was adequate.		
8	There was an equal opportunity for everyone in my group to contribute his/her ideas and opinions.		
9	I was able to follow the instructions and complete the rating sheets accurately.		
10	The discussions after the first round of ratings were helpful to me.		
11	The discussions after the second round of ratings were helpful to me		
12	The information showing the distribution of student scores was helpful to me.		
13	I am confident about the defensibility and appropriateness of the final recommended cut scores.		
14	The facilities and food service helped create a productive and efficient working environment.		

15) Comments:_____

_____

_____

FIGURE 3. Sample evaluation form for standard-setting participants.

standard-setting meetings can be long and arduous activities, it is considered desirable to conduct the final evaluation in such a way as to make the task relatively easy for participants to complete and to lessen the proportion of nonresponse. Consequently, open-ended survey items requiring lengthy responses are generally avoided. One simple modification of the evaluation form shown in Figure 3 would be to replace the Agree/Disagree options with a Likert-type scale that gives participants greater response options (e.g., 1 = Strongly Disagree to 4 = Strongly Agree). Such a modification would permit finer grained reporting of participants' perceptions, calculations of means and standard deviation for each question on the survey, and so on.

These activities all focus on an evaluation of the process. What of the product(s) of standard setting? Commonly employed criteria here include reasonableness and replicability. A first potential aspect of product evaluation is the usefulness of the PLLs and PLDs. For a given subject and grade level, they should accurately reflect the content standards or credentialing objectives

Table 7. Criteria for Evaluating Standard-Setting Procedures

Evaluation Criterion	Description
Procedural	
Explicitness	The degree to which the standard-setting purposes and processes were clearly and explicitly articulated a priori
Practicability	The ease of implementation of the procedures and data analysis; the degree to which procedures are credible and interpretable to relevant audiences
Implementation	The degree to which the following procedures were reasonable, and systematically and rigorously conducted: selection and training of participants, definition of the performance standard, and data collection
Feedback	The extent to which participants have confidence in the process and in resulting cut score(s)
Documentation	The extent to which features of the study are reviewed and documented for evaluation and communication purposes
Internal	
Consistency within method	The precision of the estimate of the cut score(s)
Intrapanelist consistency	The degree to which a participant is able to provide ratings that are consistent with the empirical item difficulties, and the degree to which ratings change across rounds
Interpanelist consistency	The consistency of item ratings and cut scores across participants
Decision consistency	The extent to which repeated application of the identified cut scores(s) would yield consistent classifications of examinees
Other measures	The consistency of cut scores across item types, content areas, and cognitive processes
External	
Comparisons to other standard-setting methods	The consistency of cut scores across replications using other standard-setting methods
Comparisons to other sources of information	The relationship between decisions made using the test to other relevant criteria (e.g., grades, performance on tests measuring similar constructs, etc.)
Reasonableness of cut scores	The extent to which cut score recommendations are feasible or realistic (including pass/fail rates and differential impact on relevant subgroups)

Source: Adapted from Pitoniak (2003).

and be reasonably consistent with statements developed by others with similar goals.

Reasonableness can be assessed by the degree to which cut scores derived from the standard-setting process being evaluated classify examinees into groups in a manner consistent with other information about the examinees. For example, suppose it could be assumed that a state's eighth-grade reading test and the NAEP were based on common content standards (or similar content standards that had roughly equal instructional emphasis). In such a case, a standard-setting procedure for the state test resulting in 72% of the state's eighth graders being classified as *Proficient*, while NAEP results for the same grade showed that only 39% were *Proficient*, would cause concern that one or the other set of standards was inappropriate.

Local information can also provide criteria by which to judge reasonableness. Do students who typically do well in class and on assignments mostly meet the top standard set for the test, while students who struggle fall into the lower categories? In the end, regardless of how reasonable a set of performance standards seems to assessment professionals or those who participated in the actual standard-setting activity, those standards will need to be locally reproducible—at least in an informal sense— in order to be widely accepted and recognized.

Replicability is another possible avenue for evaluating standard setting. For example, in some contexts where great resources are available, it is possible to conduct independent applications of a standard-setting process to assess the degree to which independent replications yield similar results. Evaluation might also involve comparisons between results obtained using one method and an independent application of one or more different methods. Interpretation of the results of these comparisons, however, is far from clear. For example, Jaeger (1989) has noted that different methods will yield different results, and there is no way to determine that one method or the other produced the wrong results. Zieky (2001) noted that there is still no consensus as to which standard-setting method is most defensible in a given situation. Again, differences in results from two different procedures would not be an indication that one was right and the other wrong; even if two methods did produce the same or similar cut scores, we could only be sure of precision, not accuracy.

The aspects of standard-setting evaluation listed here do not cover all of the critical elements of standard setting that can yield evidence about the soundness of a particular application. The preceding paragraphs have only attempted to highlight the depth and complexity of that important task. Table 7 provides a more inclusive list and description of evaluation criteria that can be used as sources of evidence bearing on the quality of the standard-setting process.

Conclusion

Setting performance standards has been called "the most controversial problem in educational assessment today" (Hambleton, 1998, p. 103). As long as important decisions must be

made, and as long as test performance plays a part in those decisions, it is likely that controversy will remain. At least to some degree, however, any controversy can be minimized by crafting well conceived methods for setting performance standards, implementing those methods faithfully, and gathering sound evidence regarding the validity of the process and the result.

Notes

[1]Some sources refer to participants in standard-setting procedures as *judges*.

[2]It should be noted that this definition addresses only one aspect of the legal theory known as due process. According to the legal theory, governmental actions concerning a person's life, liberty, or property must involve due process—that is, a systematic, open process, stated in advance, and applied uniformly. The theory further divides the concept of due process into *procedural due process* and *substantive due process*. Whereas procedural due process provides guidance regarding what elements of a procedure are necessary, substantive due process characterizes the *result* of the procedure. The notion of substantive due process demands that the procedure lead to a decision that is fundamentally fair. Whereas Cizek's definition clearly sets forth a procedural conception of standard setting, it fails to address the result of standard setting. This aspect of fundamental fairness is similar to what has been called the "consequential basis of test use" (Messick, 1989, p. 84).

[3]Though describing all procedures is beyond the scope of this module, it should be noted that participants were prepared and facilitation of this standard setting followed standard practice as regards advance materials provided to participants, orientation and training, monitoring of the process, and so on.

[4]We note one difference between the preceding formulation and that presented in Wright and Stone (1979). While Wright and Stone use β to represent examinee ability, we have used θ here and in the rest of this discussion for the sake of consistency with Equations 1–4.

[5]As an anonymous reviewer of this manuscript pointed out, the simplicity of judgment comes at a cost, which is the potential for either positive or negative bias depending on the characteristics of the test. The potential for bias arises because the method is based on an implicit judgment of whether the probability of correct response at the cut score is greater than .5. To illustrate, suppose that a test were composed of identical items that all had a probability of correct response at the cut score of .7. A participant should judge that the borderline examinee will answer all items correctly, and the resulting performance standard would be a perfect score.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *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.

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

Chinn, R. N., & Hertz, N. R. (2002). Alternative approaches to standard-setting for licensing and certification examinations. *Applied Measurement in Education, 15,* 1–14.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93–106.

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

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

Cizek, G. J. (2001a). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27.

Cizek, G. J. (2001b). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Mahwah, NJ: Erlbaum.

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

Ferrara, S., Perie, M., & Johnson, E. (2002, April). *Setting performance standards: The item descriptor (ID) matching procedure.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hambleton, R. M. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. N. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 87–114). Washington, DC: Council of Chief State School Officers.

Hambleton, R. M., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8,* 41–56.

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, CA: Jossey-Bass.

Huynh, H. (2000, April). *On item mappings and statistical rules for selecting binary items for criterion-referenced interpretation and Bookmark standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34,* 353–366.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35* (1), 69–81.

Individuals with Disabilities Education Act. (1997). Public Law 105-17 (20 U.S.C. 1412a, 16–17).

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

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

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Standard performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Erlbaum.

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

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring.* Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

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

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

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175–218). Mahwah, NJ: Erlbaum.

Mehrens, W. A., & Cizek, G. J. (2001). Standard setting and the public good: Benefits accrued and anticipated. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and per-*

spectives (pp. 477–485). Mahwah, NJ: Erlbaum.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.
Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.
Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3–19.
No Child Left Behind Act. (2001). Public Law 107–110 (20 U.S.C. 6311).
Pitoniak, M. J. (2003). *Standard setting methods for complex licensure examinations.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.
Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283–312). Mahwah, NJ: Erlbaum.
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.
Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Erlbaum.
Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task. The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–174). Mahwah, NJ: Erlbaum.
Schagen, I., & Bradshaw, J. (2003, September). *Modeling item difficulty for Bookmark standard setting.* Paper presented at the annual meeting of the British Educational Research Association, Edinburgh.
Talente, G., Haist, S., & Wilson, J. (2003). A model for setting performance standards for standardized patient examinations. *Evaluation and the Health Professions, 26*(4), 427–446.
Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement, 40,* 231–253.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA.
Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA.
Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–52). Mahwah, NJ: Erlbaum.

Annotated Bibliography

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

This ITEMS module is the precursor to the current module. It describes standard setting for achievement measures, with a focus on methods applied in the context of selected-response item formats. In addition to description of specific methods, it provides background and context for standard setting and describes issues surrounding standard setting.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Erlbaum.

This fairly recent volume contains chapters written by some of the most authoritative and experienced persons working in the field of standard setting. The book covers all aspects of standard setting, including theoretical foundations, methodologies, and current perspectives on legal issues, validation, social significance, and applications for special populations and computer-adaptive testing.

Hansche, L. N. (Ed.). (1988). *Handbook for the development of performance standards.* Washington, DC: Council of Chief State School Officers.

This handbook focuses on methods for developing performance standards in the aligned system of standards and assessments required by IASA/Title I. Sections 1 and 2 provide definitions of performance standards in the context of an aligned educational system, advice for those developing systems of performance standards, and information about experiences of several states regarding standards-based assessment systems. Section 3 contains reports about research on developing performance standards and setting cut scores on complex performance assessments.

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

This chapter appears in the foundational reference text for the field of educational measurement. One of 18 chapters, this chapter provides an overview of standard-setting methods, issues, and concerns for the future. A revision of this chapter, focusing exclusively on standard setting and to be written by R. M. Hambleton and M. J. Pitoniak will be included in the forthcoming 4th edition of *Educational Measurement.*

Self-Test

Multiple-Choice Items

1. The *Standards for Educational and Psychological Testing* (1999) require all of the following related to standard setting *except:*
   A. estimates of classification/decision consistency.
   B. description of the qualifications and experience of participants.
   C. scientifically based (i.e., experimental) standard-setting study designs.
   D. estimates of standard errors of measurement for scores in the regions(s) of recommended cut scores.
2. The typical role of the standard-setting panel is to
   A. determine one or more cut scores for a particular test.
   B. recommend one or more cut scores to authorized decision makers.
   C. determine the most appropriate method to use for the standard-setting task.
   D. develop performance level descriptors that best match the target examinees.
3. *Performance standard* is to *passing score* as
   A. *practical* is to *ideal.*
   B. *decision* is to *process.*
   C. *objective* is to *subjective.*
   D. *conceptual* is to *operational.*
4. *Performance level label* (PLL) is to *performance level descriptor* (PLD) as *title* is to
   A. index.
   B. summary.
   C. main idea.
   D. first draft.
5. Which of the following is an example of a performance standard?
   A. Students should be able to apply enabling strategies and skills to learn to read and write including inferring word meanings from taught roots, prefixes, and suffixes to de-

code words in text to assist in comprehension.

B. To be prepared for the jobs of the future, students must demonstrate an understanding of the overall meaning of what they read. When reading appropriate grade-level text, they should be able to make relatively obvious connections between the text and their own experiences and extend the ideas in the text by making simple inferences.

C. To be considered "Accelerated" students must obtain at least 35 points on the set of seven constructed-response items designed to assess grade-level reading comprehension.

D. Students performing at the "Accelerated" level consistently demonstrate mastery of grade-level subject matter and skills and are well prepared for the next grade level.

6. Which of the following is true regarding the composition of a standard-setting panel?
   A. It should consist of at least 10 members for each construct measured by a multidimensional test.
   B. It should include only participants with previous standard-setting experience.
   C. It should be diverse enough to represent all likely examinee demographics.
   D. It should be large and representative enough to produce reliable results.
7. A primary benefit of the Yes/No method is that it
   A. simplifies the decision-making task for participants.
   B. increases the sensitivity of participants to impact data.
   C. reduces the need for participants to be familiar with typical examinee ability.
   D. increases the likelihood that the true cut score will result from the standard-setting process.
8. Suppose that a decision was made to require examinees to obtain a score of at least 2 on each of the CR items shown in Table 6 (and that the ratings and decision rules for the SR items remained the same). What would be the result on the cut score for the total test?
   A. The cut score would remain the same.
   B. The cut score would increase by about 5 raw score points.
   C. The cut score would decrease by about 5 raw score points.
   D. Cannot determine; the result would depend on examinee performance on the CR items.
9. Suppose that a sixth-grade reading test consisted of four reading passages, each of which was followed by eight multiple-choice items and one constructed-response item. Using the Bookmark method, which would be the most appropriate way to construct an ordered item booklet for this test?
   A. First arrange the passages in increasing order of readability, then arrange the items for each passage in order of increasing difficulty.
   B. Arrange all items in difficulty order, printing the appropriate portion of the passage on the individual item pages.
   C. Arrange all items in difficulty order, with reference to the appropriate passage on each page, printing all passages in a separate booklet.
   D. Arrange the test booklet to be identical to the one students used, printing at the top of each page the difficulty index of the item and its rank order.
10. Suppose that facilitators for a standard-setting study using a Bookmark method trained participants to use an "RP50" decision rule to set a cut score for Proficient. In this situation, RP50 refers to the probability that
    A. 50% of examinees who answer this item correctly will be considered Proficient.
    B. 50% of borderline-Proficient examinees will answer this item correctly.
    C. 50% of all Proficient examinees will answer this item correctly.
    D. 50% of all examinees will answer this item correctly.
11. Suppose that standard-setting participants have completed their Round 1 ratings for Basic using the Bookmark method. Which of the following pieces of data would be used to calculate the Round 1 cut score for Basic?
    A. Page number only
    B. Page number and item difficulty
    C. Student ability (theta) estimate
    D. Standard deviation of the Round 1 ratings
12. Which of the following scenarios would most likely be classified as a "holistic" standard-setting procedure?
    A. Standard setters review standardized math portfolios produced by 35 different students.
    B. Standard setters review sample performances by 200 students on a single writing prompt.
    C. Standard setters estimate the likelihood of a minimally Proficient student answering each of 60 multiple-choice items correctly.
    D. Standard setters compare the performances of a group of known experts in a field with the performances of a group of known novices.
13. Which information would most likely be withheld from standard-setting participants during the second round of a holistic standard-setting activity?
    A. performance level descriptors
    B. individual student scores on the tests
    C. distributions of student scores on the tests
    D. cut scores from the earlier round of judgments

To answer Item 14, refer to Figure 2 in the Module.

14. In Figure 2, what is the rationale for setting a cut score at 39 points?
    A. A score of 39 points treats misclassifications of *Proficient* and *Advanced* as equally serious.
    B. Fifty percent of the examinees in the *Advanced* group had raw scores of 39 or higher.

C. The midpoint between the means of work samples rated as *Proficient* and *Advanced* is 39.

D. The mean score for work samples rated as *Advanced* in Round 1 is 39.

Constructed-Response Item

15. Develop one additional survey item that would be appropriate for inclusion in the list of evaluation items shown in Figure 3.

Answer Key to Self-Test

1. C
2. B
3. D
4. B
5. C
6. D
7. A
8. C
9. C
10. B
11. C
12. A
13. B
14. A
15. Answers will vary, but may include items such as:

    "The members of my group brought diverse perspectives to the discussions."

    "I felt qualified to make the judgments we were asked to make."

    "The data we received showing probable effects of our ratings on pass/fail rates was a helpful piece of information."

    "Reviewing the content standards that were sent prior to the meeting helped me understand the purpose of the test."