10

The Bookmark Method

The Bookmark procedure may be viewed as a logical successor to a series of item-mapping strategies developed in the 1990s in conjunction with standard settings carried out for the National Assessment of Educational Progress (NAEP) by researchers at American College Testing (ACT). Early item-mapping techniques were applied less as standard-setting procedures per se than as feedback mechanisms embedded in other procedures (cf. Loomis & Bourque, 2001).

In 1996, for example, researchers at ACT employed an item-mapping procedure in conjunction with a method they referred to as Mean Estimation, which was essentially an extension of the modified Angoff (1971) technique. That item-mapping procedure was applied to tests with both multiple-choice and constructed-response items (Loomis, Bay, Yang, & Hanick, 1999). Item maps were used to provide feedback after a second round of item ratings for the 1996 Science assessment and the 1998 NAEP Civics and Writing assessments. The maps showed the location of each item in relation to the NAEP-like scale score, which was also associated with the various NAEP achievement level descriptors (ALDs, which are now commonly referred to as performance level descriptors). Each multiple-choice item was mapped in accordance with its probability of correct response for each scale score, and each constructed-response item was mapped once for each score point, that is, for the probability of obtaining a score of 1, 2, 3, or higher at each scale score point.

Item-mapping techniques evolved through the course of several NAEP standard-setting studies at ACT. The Reckase chart (Reckase, 2001) was introduced as a way to simplify the task set before participants. With

Reckase charts, participants would receive their Round 2 item estimates (i.e., the probability of a correct response by a student at the cut score for multiple-choice items and estimated raw score for constructed-response items for this same student or group of students), along with a preprinted table or "map" of item probabilities.

A sample Reckase chart for an individual participant is shown in Table 10-1. A unique Reckase chart would be developed based on each participant's item ratings. The first column in the Reckase chart shown in the table presents scaled scores arranged from high to low. Scaled scores are used in Reckase charts as a measure of overall examinee competence or ability on whatever construct is measured by the test. Each of the remaining columns contains information on a single item. Table 10-1 shows information on five items with Items 1–4 being dichotomously scored multiple-choice format items and Item 5 being a constructed-response item scored on a 0–5 scale. For the multiple-choice items, the data in each column show the probability of an examinee at each scaled score answering that item correctly, based on the three-parameter item response model. For example, an examinee with an overall ability level (i.e., scaled score) of 170 has a .53 probability of answering Item 1 correctly. For constructed-response items, the values in a column show the expected item score for examinees at a given scaled score location. Again considering an examinee with an ability level of 170, the expected score of that examinee on the constructed-response item (Item 5) is 1.8 out of 5.

In Table 10-1, one value in each column appears in brackets; it is in this way that Reckase charts are individualized for each participant. When used as feedback in standard setting, Reckase charts help participants gauge how consistently they are applying their conceptualization of the minimally competent examinee, borderline candidate, or whatever hypothetical examinee is considered. The Reckase chart for a participant who is consistently applying his or her conceptualization would show brackets aligned in a single row. For example, consider the participant whose judgments resulted in the values shown in the table. In addition, let us assume that the participant held an implicit conceptualization that the minimally qualified examinee is one with an ability level (represented by a scaled score) of 170. Reading across the row in the table corresponding to a scaled score of 170, we see that the probability estimate (i.e., Angoff rating) generated by this participant was .53; this participant is saying that the probability of a minimally qualified examinee answering Item 1 correctly is .53. Now, if this participant were applying his or her conceptualization of the minimally qualified examinee consistently, he or she would have generated an Angoff rating of .83 for Item 2, .34 for Item 3, and .77 for Item 4. For the

constructed-response item (Item 5), this participant would have estimated the minimally qualified examinee's score to be 1.8 out of 5.

From the Reckase chart shown in Table 10-1, however, the participant can see that he or she is not making totally consistent judgments. For the remaining three multiple-choice items (Items 2–4), the participant has estimated the items to be more difficult than they are for an examinee of ability level 170. For example, for Item 2, the participant judged the minimally qualified examinee to have a .57 probability of success on the item when, using the standard implied by this participant's rating of Item 2, the rating for Item 2 should have been .83. For the constructed-response item, the reviewer exhibited more consistent behavior with his or her implicit performance standard as shown by the fact that his or her rating of Item 5 of 1.5 is very close to the expected constructed-response item score of 1.8 for examinees with an overall ability level of 170. If this participant were being perfectly consistent, the bracketed values would be aligned in a row corresponding to a single ability level (scaled score).

Table 10-1 can be thought of as an early item map. From this foundation, it was not a great step to refine the item-mapping procedure by reordering the items according to their difficulty. Loomis, Hanick, Bay, and Crouse (2000) reported on field trials for the 1998 NAEP Civics test in which the item maps were reordered from least to most difficult item. These item maps also included brief descriptions of item content, which permitted participants, at a glance, to summarize both the location and content of an item and to reframe their own judgments of those items. From difficulty-ordered item maps with content information and probability of correct response, the leap to an ordered test booklet with similar information was a short but significant one. Researchers at CTB/McGraw-Hill made that leap and introduced the Bookmark method (Lewis, Mitzel, & Green, 1996).

Overview of the Bookmark Method

The standard Bookmark procedure (Mitzel et al., 2001) is a complete set of activities designed to yield cut scores on the basis of participants' reviews of collections of test items. The Bookmark procedure is so named because participants express their judgments by entering markers in a specially designed booklet consisting of a set of items placed in difficulty order, with items ordered from easiest to hardest. This booklet, called an *ordered item booklet,* will be described in greater detail in the next portion of this chapter.

The Bookmark procedure has become quite popular for several reasons. First, from a practical perspective, the method can be used for complex,

Table 10-1 Example of a Reckase Chart

	Probabilities of Correct Response for Given Scale Score				
Scale Score	*Item 1*	*Item 2*	*Item 3*	*Item 4*	*Item 5*
215	.99	.99	.99	.99	4.8
212	.99	.99	.98	.99	4.7
209	.99	.99	.97	.98	4.6
206	.98	.99	.96	.98	4.5
203	.98	.99	.94	.97	4.4
200	.97	.99	.91	.97	4.3
197	.96	.98	.88	.96	4.1
194	.94	.98	.83	.95	3.9
191	.92	.97	.77	.94	3.7
188	.89	.96	.70	.92	3.5
185	.85	.95	.63	.91	3.2
182	.81	.93	.55	.89	2.9
179	.75	.91	.48	.86	2.6
176	.68	.89	.42	.83	2.4
173	.60	.86	.37	.80	2.1
170	[.53]	.83	.34	.77	1.8
167	.45	.79	.31	.73	[1.5]
164	.37	.74	.29	.69	1.3
161	.30	.69	.28	.66	1.1
158	.24	.63	.27	.62	0.9
155	.20	[.57]	.26	.58	0.7
152	.16	.52	.26	.55	0.6
149	.13	.46	.26	.52	0.5
146	.11	.41	.25	.49	0.4
143	.09	.36	.25	.46	0.3
140	.08	.32	.25	[.44]	0.2
137	.07	.29	.25	.43	0.2
134	.07	.26	.25	.41	0.2
131	.06	.24	.25	.40	0.1
128	.06	.22	.25	.39	0.1
125	.06	.20	[.25]	.38	0.1

NOTES: For multiple-choice items (Items 1–4) the values in brackets [ ] are a participant's Angoff ratings; for constructed-response items (Item 5) the value in brackets is the participant's estimated mean score for a minimally competent examinee.

Source: Adapted from Reckase (2001).

mixed-format assessments, and participants using the method consider selected-response (SR) and constructed-response (CR) items together. As the prevalence of mixed-format examinations continues to increase, it is likely that the Bookmark method will become even more widely used and that other innovative approaches for setting performance standards in such contexts will be developed.

Second, from the perspective of those who will be asked to make judgments via this method, it presents a relatively simple task to participants, and one with which, at a conceptual level, they may already be familiar. To fully grasp the extent to which the Bookmark method simplifies the standard-setting task, it is instructive to consider a test with four performance levels (*Below Basic, Basic, Proficient,* and *Advanced*), 60 SR items, and four CR items (with four score points each). If item-based standard-setting methods such as the Angoff or modified Angoff procedures were used, participants would have 192 separate tasks to perform per round of ratings (i.e., three probability judgments for each of 64 items). With the Bookmark procedure, the same participant may still consider the content covered by the items in a test but is required to make only three judgments—one for each of three bookmarks (*Basic, Proficient,* and *Advanced*) he or she will be asked to place in a difficulty-ordered test booklet (described in more detail later in this chapter). The task is perhaps even more streamlined because it would seem reasonable that the bookmark for *Advanced* should be placed after the bookmark for *Proficient,* and that the bookmark for *Proficient* should be after the bookmark for *Basic*. Thus once a participant has identified one cut score through the placement of his or her bookmark, it is not necessary for him or her to start the search for the next cut score at the beginning of the ordered test booklet. In order to make judgments about each subsequent cut score, participants can examine a relatively narrow range of items rather than reexamining each item and making a new estimate of the probability of a student just barely at a particular performance level answering correctly.

Third, in addition to being relatively easy for participants, the Bookmark method is also comparatively easy for those who must implement the procedure. Although some of the computational aspects of the method are mathematically complex, most of the intensive work is done long before the standard-setting session itself occurs. For those who conduct such sessions, this is an important feature of the procedure that helps reduce the potential for errors and the time required for the standard-setting meeting.

Finally, from a psychometric perspective, the method has certain advantages because of its basis in item response theory (IRT) analyses, and because of the fidelity of the method to the test construction techniques that spawned the assessment. With few exceptions, most high-stakes, large-scale

assessments are constructed in accordance with an IRT model, either Rasch or 3PL. The analyses normally carried out in the construction and equating of these tests make IRT-based standard-setting procedures a natural extension. Once participants provide page numbers, the associated theta values have a built-in relationship to scores, and results can be interpreted in the same manner as other procedures carried out with these tests. In the absence of other IRT-based standard-setting procedures, the bookmark procedure is a natural choice.

The Ordered Item Booklet

Perhaps the most distinctive feature of the Bookmark method is the collection of items that serves as the focus of participants' judgments. This booklet, called an *ordered item booklet* (OIB), can contain both SR format items, such as multiple-choice, and CR items intermingled in the same booklet. An SR item appears in the OIB once, in a location determined by its difficulty (usually its IRT *b* value). Each CR item appears several times in the booklet, once for each of its score points. For a typical application of the procedure, each SR item will have one associated difficulty index, and each CR item will have as many step (difficulty) functions as it has score points (excluding zero). For a given CR entry, the item prompt and the rubric for a particular score point would ordinarily also be provided to participants, along with sample responses illustrating that score point.

The OIB can be composed of any collection of items spanning the range of content, item types, and difficulty represented in a typical test and need not consist only of items that have appeared in an intact test. This booklet can have more items or fewer items than an operational test form. One advantage of permitting items beyond those included in an operational test form is the fact that gaps in item difficulty or content coverage can be filled with items from a bank. For example, if two adjacent items in the ordered booklet have difficulty indices of 1.05 and 1.25 logits, additional items with difficulty indices of 1.10, 1.15, and 1.20 could be inserted to help standard-setting participants place their bookmarks more precisely. Conversely, a clear advantage of using an intact test form for standard setting using the Bookmark method is the fact that the results can be interpreted in a straightforward manner; namely, the test booklet on which standards are set is the same set of items on which student scores and (sometimes high-stakes) decisions are based.

Figure 10-1 shows the general layout of a hypothetical OIB. As indicated previously, items in the OIB appear one per page. Each SR item appears on a single page; each CR item is included in the OIB a number of times equal to the number of possible score points (excluding zero) associated with the item, along with one or more sample responses at that score point. Thus an

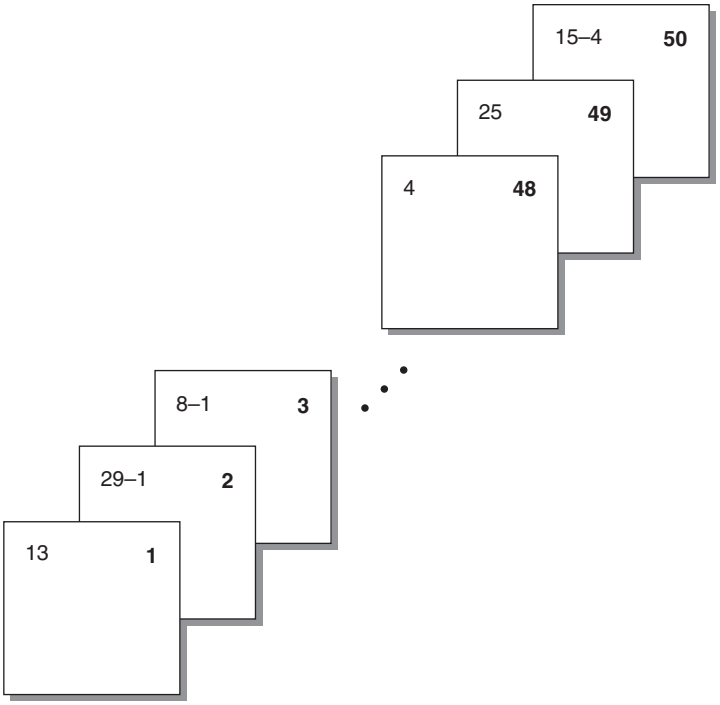

Figure 10-1 Hypothetical Illustration of an Ordered Item Booklet

item scored on a 5-point scale (0–4) and thus having four nonzero score points (i.e., 1, 2, 3, 4) would be represented on four different pages in the OIB. These configurations are shown in Figure 10-1. The bold numbers at the top right of each page illustrated in the figure simply indicate the sequence of the items in the OIB (i.e., pagination). The numbers at the top left indicate the positions in the test form upon which the OIB is based. For example, the item appearing on the first page of the OIB appeared as Item 13—an SR item—in the intact test form. It should be noticed that some of these numbers in the top left corner of the OIB pages have hyphens. These numbers refer to the original item number and the score point represented on that page. For example, the second page in the OIB represents a response earning a score of 1 to original Item 29 (a CR item); page 50 in the OIB contains the response earning the highest score (i.e., 4) on another CR item (Item 15 in the original intact test form). In an actual OIB, information beyond simple pagination and original item numbers would be included. A more detailed description and illustration of the information typically provided on an OIB page is presented later in this chapter.

The Response Probability (RP) Value

In the Bookmark procedure, the basic question participants must answer is "Is it likely that the minimally qualified or borderline examinee will answer this SR item correctly (or earn this CR item score point)?" Obviously, it is important to define "likely" or to operationalize this decision rule. In practice, the Bookmark procedure employs a 67% likelihood (or sometimes a 2/3 chance) of desired response (i.e., of getting the SR item correct or of achieving a certain CR score point or higher).

In the more than 30 years that have intervened between the introduction of the Angoff and Bookmark methods, there has been considerable experimentation with decision rules. Huynh (2000, 2006) has argued that the probability value that maximizes the information of the correct response would produce the optimum decision rule. As it turns out, for a three-parameter model with the guessing parameter removed (i.e., a two-parameter model), a 67% likelihood (i.e., a response probability [RP] of .67) optimizes this value. Thus the typical decision rule for the bookmark procedure is .67, although other percentages (ranging from .50 to .80) are also sometimes used.

In a Rasch model context, Wang (2003) has expressed a preference for a 50% likelihood (RP = .50). Indeed, the choice of .50 for the Rasch model has certain practical advantages over .67 in that the likelihood of a correct response is exactly .50 when the examinee's ability is equal to the item's difficulty. Wang pointed out, however, that the issue should not be considered resolved and urged further research into the efficacy of the .50 decision rule in Rasch applications. Although the difference may at first seem trivial, following both the suggestion of the originators of the Bookmark procedure and our own experience in implementing the Bookmark method, our tendency is to use a decision rule of 2/3. We note too that we tend to express the decision rule in this way (rather than as RP = .67). Of course, framing the issue as a decision rule of 2/3 or as an RP of .67 is (at least mathematically) nearly the same. In our experience, however, standard-setting participants seem better able to grasp and work with the notion of "two out of three" more readily than a probability of .67.

Response Probabilities and Ordered Item Booklet Assembly—Rasch Model

As may already be obvious, the choice of a decision rule (or RP value) is essential to the assembly of OIBs and to the calculation of cut scores when the Bookmark method is used. In the following description, we assume that a Rasch model has been used for test construction, item calibration, and so

on and that a decision rule of 2/3 has been incorporated into participants' training, practice, and OIB rating activities. We begin with the basic Rasch equation, set forth in Wright and Stone (1979), which expresses the probability of answering an item correctly, $p(x = 1)$, as a function of the item's difficulty (β_j) and the examinee's ability (θ_i):

$$p(x = 1|\theta_i, \beta_j) = \exp(\theta_i - \beta_j)/[1 + \exp(\theta_i - \beta_j)] \quad \text{(Equation 10-1)}$$

Now, setting p equal to 2/3 and solving for θ_i we obtain

$$\exp(\theta_i - \beta_j)/[1 + \exp(\theta_i - \beta_j)] = 2/3 \quad \text{(Equation 10-2)}$$

$$\exp(\theta_i - \beta_j) = 2/3 * [1 + \exp(\theta_i - \beta_j)] \quad \text{(Equation 10-3)}$$

$$\exp(\theta_i - \beta_j) = 2/3 + 2/3 * \exp(\theta_i - \beta_j) \quad \text{(Equation 10-4)}$$

$$\exp(\theta_i - \beta_j) - 2/3 * \exp(\theta_i - \beta_j) = 2/3 \quad \text{(Equation 10-5)}$$

$$1/3 * \exp(\theta_i - \beta_j) = 2/3 \quad \text{(Equation 10-6)}$$

$$\exp(\theta_i - \beta_j) = 2/3 \div 1/3 \quad \text{(Equation 10-7)}$$

$$\exp(\theta_i - \beta_j) = 2 \quad \text{(Equation 10-8)}$$

Finally, taking the natural log of both sides of Equation 10–8, we obtain

$$\theta_i - \beta_j = .693, \text{ and} \quad \text{(Equation 10-9)}$$

$$\theta_i = \beta_j + .693 \quad \text{(Equation 10-10)}$$

The reader who is familiar with the work of Wright and Stone (1979) will notice that we have used slightly different notation than that source. Our substitution of θ and β to represent examinee ability and item difficulty, respectively, is an attempt to make the notation used in the preceding explication more consistent with standard notion across the family of IRT models. (We should also note another small but important difference between a 2/3 decision rule and an RP67 rule. If a response probability of .67 had been used, Equation 10-10 would have been $\theta_i = \beta_j + .708$; that is, it would have been computed by taking the item's difficulty plus the natural logarithm of .67/.33.)

The use of the final result in Equation 10-10 to assemble the OIB is straightforward. For SR items, to calculate the value of θ_i needed to have

a 2/3 chance of answering a given SR item correctly, we simply add .693 to the Rasch difficulty value for that item, where the Rasch difficulty of an item is obtained by use of an IRT calibration program (e.g., WINSTEPS). As is perhaps evident, when the Rasch model is used to create an OIB with SR items, the procedure just described will result in the same ordering of items in the OIB as if the booklet had been assembled using the items' *b* values. This result would *not* likely occur, however, for items calibrated using a 2PL or 3PL model. As we will see a bit later, these same values used to determine the placement of SR items in the difficulty-ordered test booklet are also used in determining the raw score associated with setting a bookmark right after this item in the OIB.

Locating the appropriate placement of CR items in the OIB is only slightly more complicated. To locate the score points of CR items in the OIB within a Rasch framework, the Partial-Credit Model (PCM; Wright & Masters, 1982) is used. In the following discussion, the procedure is illustrated for a CR item with five score points (0, 1, 2, 3, 4); however, the logic is applied to items with any number of steps.

To begin, for CR items, the likelihood (π_{nix}) of a person with a given ability (θ_n) obtaining any given score (j) in any item (i) is shown in the following equation, taken from Wright and Masters (1982, equation 3.1.6):

$$\pi_{nix} = \frac{\exp \Sigma(\theta_n - \delta_{ij})}{\Sigma \exp \Sigma(\theta_n - \delta_{ij})} \qquad \text{(Equation 10-11)}$$

In Wright and Masters's formulation, the difficulties associated with each score point are referred to as step functions and are symbolized generally as δ_{ij}. The step function for score point 0 is set equal to 0 in Equation 10-11; that is, $\delta_{i0} \equiv 0$, such that

$$\Sigma(\theta_n - \delta_{ij}) = 0, \text{ and } \exp \Sigma(\theta_n - \delta_{ij}) = 1 \qquad \text{(Equation 10-12)}$$

The numerator values for the other steps are derived as follows:

Step 1. $\Sigma(\theta_n - \delta_{ij}) = \Sigma(\theta_n - \delta_{i0}) + \theta_n - \delta_{i1}$
$= 0 + \theta_n - \delta_{i1}$
$= \theta_n - \delta_{i1}$ (Equation 10-13)

Step 2. By similar logic: $\Sigma(\theta_n - \delta_{ij}) = 2\theta_n - \delta_{i1} - \delta_{i2}$ (Equation 10-14)

Step 3. By similar logic: $\Sigma(\theta_n - \delta_{ij})$
$= 3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3}$, and (Equation 10-15)

Step 4. By similar logic: $\Sigma(\theta_n - \delta_{ij})$
$= 4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4}$ (Equation 10-16)

The exponential values of these summations shown in Equations 10-12 through 10-16 are simply the natural logarithm *e* raised to the respective values, that is:

Step 0. $\exp(0)$ (Equation 10-17)

Step 1. $\exp(\theta_n - \delta_{i1})$ (Equation 10-18)

Step 2. $\exp(2\theta_n - \delta_{i1} - \delta_{i2})$ (Equation 10-19)

Step 3. $\exp(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})$ (Equation 10-20)

Step 4. $\exp(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})$ (Equation 10-21)

The denominator of Equation 10-11 now becomes the simple sum of the values yielded by Equations 10-17 through 10-21 for Steps 0–4. Finally, the desired end—the likelihood of obtaining any given score (0 through 4)—is calculated by dividing the numerator associated with that score point by this common denominator. These calculations can be carried out by hand or with various software programs, such as SPSS, SAS, or Excel. A step-by-step procedure for using Excel to compute the appropriate OIB locations for CR item score points is provided in Table 10-2. The Excel spreadsheet on which the table is based is available with the other electronic materials accompanying this book at www.sagepub.com/cizek/bookmark.

As an initial check on the accuracy of the values obtained, begin by locating the values in columns N–Q. Read down the column of values until .5000 or the closest value to .5000 is found. Then simply read across the row from this value to column A to find the corresponding value of θ_n. This value of θ_n should correspond to the Thurstone Threshold Value reported in WINSTEPS. Having verified that the RP50 value corresponds to the Thurstone Threshold Value, continue down columns N–Q (depending on the score point desired) to find the value closest to .6667, or use interpolation to obtain an exact value. Again, read across the row to column A to find the corresponding value of θ. This value is the ability (or achievement) level associated with a 2/3 chance of obtaining the particular score point or better on the CR item. These values are then used to determine the placement of the score points for CR items in the OIB and in calculating raw scores associated with setting a bookmark right after this item/score point.

Table 10-2 Excel Instructions for Calculating Ability (θ_n) for a Specified Response Probability (RP)

Column	*Excel Code/Instructions [Explanation]*
A	Enter values of θ from –4 to +4 in increments of .01 (i.e., –4.00, –3.99, –3.98, etc.).
B	Enter 1 in every row. [Numerator value for Step 0.]
C	= exp(value in Col. A – δ_{i1}). [Numerator value for Step 1.] Copy to remaining rows in this column.
D	= exp (2 * value in col. A – δ_{i1} – δ_{i2}). [Numerator value for Step 2.] Copy to remaining rows in this column.
E	= exp(3 * value in col. A – δ_{i1} – δ_{i2} – δ_{i3}). [Numerator value for Step 3.] Copy to remaining rows in this column.
F	= exp(4 * value in col. A – δ_{i1} – δ_{i2} – δ_{i3} – δ_{i4}). [Numerator value for Step 4.] Copy to remaining rows in this column.
G	= sum(values in col. B – F). [Denominator.] Copy to remaining rows in this column.
H	= (value in col. B)/(value in col. G). [Probability value for Step 0.] Copy to remaining rows in this column.
I	= (value in col. C)/(value in col. G). [Probability value for Step 1.] Copy to remaining rows in this column.
J	= (value in col. D)/(value in col. G). [Probability value for Step 2.] Copy to remaining rows in this column.
K	= (value in col. E) / (value in col. G). [Probability value for Step 3.] Copy to remaining rows in this column.
L	= (value in col. F)/(value in col. G). [Probability value for Step 4.] Copy to remaining rows in this column.
M	= sum(values in col. H – L). [Sum of the probability values.] Copy to remaining rows in this column. Note: This can be used as a check on the accuracy of calculated values. For any given value of θ_n, the sum of the probabilities should be 1.00.
N	= sum(values in col. I – L). [Probability of obtaining a score of 1 or better.] Copy to remaining rows in this column.
O	= sum(values in col. J – L). [Probability of obtaining a score of 2 or better.] Copy to remaining rows in this column.
P	= sum(values in col. K – L). [Probability of obtaining a score of 3 or better.] Copy to remaining rows in this column.
Q	= (value in col. L). [Probability of obtaining score of 4.] Copy to remaining rows in this column.

Other software programs can be used to calculate the RP50 and RP67 (or P2/3) values without displaying all the results from all the intermediate steps. In our experience, however, it is often helpful to be able to review all the intermediate values because they can be used to create item characteristic curves and to check the accuracy of results along the way. For example, as we alluded to previously, WINSTEPS produces a threshold value for each step of a CR item, which is equivalent to the RP50 value for the item. Table 10-3 shows a portion of an Excel spreadsheet for a set of calculations for a hypothetical 4-point CR item where the items for the test were scaled using the Rasch model. The step values associated with each of the four score points are provided at the bottom of the table. Figure 10-2 shows the response characteristic curves associated with each option for that item, and Figure 10-3 shows the curves associated with the probability of obtaining a given score or better on the same item.

Response Probabilities and Ordered Item Booklet Assembly—2PL Model

Mitzel et al. (2001) note that the probability of a correct response, $p(x = 1)$, to a given 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 three-parameter logistic (3PL) model:

$$p(x = 1|\theta) = c_j + (1 - c_j)/\{1 + \exp[-1.7a_j(\theta - b_j)]\} \quad \text{(Equation 10-22)}$$

where c_j is the lower asymptote or threshold value of the item (the likelihood that an extremely low-scoring student would answer correctly by guessing), a_j is the discrimination index of the item, and b_j is the difficulty of the item. In practice, Mitzel et al. (2001) and others using this model set the threshold or chance parameter (c_j) equal to zero, reducing Equation 10-22 to the following:

$$P_j(\theta) = 1/\{1 + \exp[-1.7a_j(\theta - b_j)]\} \quad \text{(Equation 10-23)}$$

or a two-parameter logistic (2PL) model.

In the procedure described by Mitzel et al. (2001), the basic standard-setting question is whether an examinee just barely qualified for a given performance level would have a 2/3 chance of answering a given SR item

Table 10-3 Selected Spreadsheet Entries and Calculations for a Hypothetical 4-point CR Item, Rasch Scaling

	Numerator					*Denom.*	*P*						*P*			
Theta	*0*	*1*	*2*	*3*	*4*	*Sum*	*0*	*1*	*2*	*3*	*4*	*Total*	*1 or Better*	*2 or Better*	*3 or Better*	*4*
–4.00	1.0000	0.0074	0.0001	0.0000	0.0000	1.0075	0.9926	0.0074	0.0001	0.0000	0.0000	1.0000	0.0074	0.0001	0.0000	0.0000
–3.99	1.0000	0.0075	0.0001	0.0000	0.0000	1.0076	0.9925	0.0075	0.0001	0.0000	0.0000	1.0000	0.0075	0.0001	0.0000	0.0000
–3.98	1.0000	0.0076	0.0001	0.0000	0.0000	1.0077	0.9924	0.0075	0.0001	0.0000	0.0000	1.0000	0.0076	0.0001	0.0000	0.0000
–3.97	1.0000	0.0077	0.0001	0.0000	0.0000	1.0077	0.9923	0.0076	0.0001	0.0000	0.0000	1.0000	0.0077	0.0001	0.0000	0.0000
–3.96	1.0000	0.0078	0.0001	0.0000	0.0000	1.0078	0.9922	0.0077	0.0001	0.0000	0.0000	1.0000	0.0078	0.0001	0.0000	0.0000
–3.95	1.0000	0.0078	0.0001	0.0000	0.0000	1.0079	0.9922	0.0078	0.0001	0.0000	0.0000	1.0000	0.0078	0.0001	0.0000	0.0000
–3.94	1.0000	0.0079	0.0001	0.0000	0.0000	1.0080	0.9921	0.0078	0.0001	0.0000	0.0000	1.0000	0.0079	0.0001	0.0000	0.0000
–3.93	1.0000	0.0080	0.0001	0.0000	0.0000	1.0081	0.9920	0.0079	0.0001	0.0000	0.0000	1.0000	0.0080	0.0001	0.0000	0.0000
–3.92	1.0000	0.0081	0.0001	0.0000	0.0000	1.0081	0.9919	0.0080	0.0001	0.0000	0.0000	1.0000	0.0081	0.0001	0.0000	0.0000
–3.91	1.0000	0.0081	0.0001	0.0000	0.0000	1.0082	0.9919	0.0081	0.0001	0.0000	0.0000	1.0000	0.0081	0.0001	0.0000	0.0000
–3.90	1.0000	0.0082	0.0001	0.0000	0.0000	1.0083	0.9918	0.0082	0.0001	0.0000	0.0000	1.0000	0.0082	0.0001	0.0000	0.0000
1.64	1.0000	2.0959	4.3929	5.1039	2.3164	14.909	0.0671	0.1406	0.2946	0.3423	0.1554	1.0000	0.9329	0.7923	**0.4977**	0.1554
1.65	1.0000	2.1170	4.4817	5.2593	2.4109	15.268	0.0655	0.1386	0.2935	0.3444	0.1579	1.0000	0.9345	0.7959	**0.5023**	0.1579
1.66	1.0000	2.1383	4.5722	5.4195	2.5093	15.639	0.0639	0.1367	0.2924	0.3465	0.1604	1.0000	0.9361	0.7993	0.5070	0.1604
1.67	1.0000	2.1598	4.6646	5.5845	2.6117	16.020	0.0624	0.1348	0.2912	0.3486	0.1630	1.0000	0.9376	0.8028	0.5116	0.1630
1.68	1.0000	2.1815	4.7588	5.7546	2.7183	16.413	0.0609	0.1329	0.2899	0.3506	0.1656	1.0000	0.9391	0.8062	0.5162	0.1656
1.69	1.0000	2.2034	4.8550	5.9299	2.8292	16.817	0.0595	0.1310	0.2887	0.3526	0.1682	1.0000	0.9405	0.8095	0.5208	0.1682

	Numerator					Denom.	P						P			
Theta	*0*	*1*	*2*	*3*	*4*	*Sum*	*0*	*1*	*2*	*3*	*4*	*Total*	*1 or Better*	*2 or Better*	*3 or Better*	*4*
1.70	1.0000	2.2255	4.9530	6.1104	2.9447	17.233	0.0580	0.1291	0.2874	0.3546	0.1709	1.0000	0.9420	0.8128	0.5254	0.1709
1.71	1.0000	2.2479	5.0531	6.2965	3.0649	17.662	0.0566	0.1273	0.2861	0.3565	0.1735	1.0000	0.9434	0.8161	0.5300	0.1735
1.72	1.0000	2.2705	5.1552	6.4883	3.1899	18.103	0.0552	0.1254	0.2848	0.3584	0.1762	1.0000	0.9448	0.8193	0.5346	0.1762
1.73	1.0000	2.2933	5.2593	6.6859	3.3201	18.558	0.0539	0.1236	0.2834	0.3603	0.1789	1.0000	0.9461	0.8225	0.5392	0.1789
1.74	1.0000	2.3164	5.3656	6.8895	3.4556	19.027	0.0526	0.1217	0.2820	0.3621	0.1816	1.0000	0.9474	0.8257	0.5437	0.1816
1.75	1.0000	2.3396	5.4739	7.0993	3.5966	19.509	0.0513	0.1199	0.2806	0.3639	0.1844	1.0000	0.9487	0.8288	0.5482	0.1844
1.76	1.0000	2.3632	5.5845	7.3155	3.7434	20.006	0.0500	0.1181	0.2791	0.3657	0.1871	1.0000	0.9500	0.8319	0.5528	0.1871
1.77	1.0000	2.3869	5.6973	7.5383	3.8962	20.518	0.0487	0.1163	0.2777	0.3674	0.1899	1.0000	0.9513	0.8349	0.5573	0.1899
1.78	1.0000	2.4109	5.8124	7.7679	4.0552	21.046	0.0475	0.1146	0.2762	0.3691	0.1927	1.0000	0.9525	0.8379	0.5618	0.1927
1.79	1.0000	2.4351	5.9299	8.0045	4.2207	21.590	0.0463	0.1128	0.2747	0.3707	0.1955	1.0000	0.9537	0.8409	0.5662	0.1955
1.80	1.0000	2.4596	6.0496	8.2482	4.3929	22.150	0.0451	0.1110	0.2731	0.3724	0.1983	1.0000	0.9549	0.8438	0.5707	0.1983
1.81	1.0000	2.4843	6.1719	8.4994	4.5722	22.727	0.0440	0.1093	0.2716	0.3740	0.2012	1.0000	0.9560	0.8467	0.5751	0.2012
1.82	1.0000	2.5093	6.2965	8.7583	4.7588	23.322	0.0429	0.1076	0.2700	0.3755	0.2040	1.0000	0.9571	0.8495	0.5796	0.2040
1.83	1.0000	2.5345	6.4237	9.0250	4.9530	23.936	0.0418	0.1059	0.2684	0.3770	0.2069	1.0000	0.9582	0.8523	0.5840	0.2069
1.84	1.0000	2.5600	6.5535	9.2999	5.1552	24.568	0.0407	0.1042	0.2667	0.3785	0.2098	1.0000	0.9593	0.8551	0.5884	0.2098
1.85	1.0000	2.5857	6.6859	9.5831	5.3656	25.220	0.0397	0.1025	0.2651	0.3800	0.2127	1.0000	0.9603	0.8578	0.5927	0.2127
1.86	1.0000	2.6117	6.8210	9.8749	5.5845	25.892	0.0386	0.1009	0.2634	0.3814	0.2157	1.0000	0.9614	0.8605	0.5971	0.2157
1.87	1.0000	2.6379	6.9588	10.175	5.8124	26.584	0.0376	0.0992	0.2618	0.3828	0.2186	1.0000	0.9624	0.8632	0.6014	0.2186
1.88	1.0000	2.6645	7.0993	10.485	6.0496	27.299	0.0366	0.0976	0.2601	0.3841	0.2216	1.0000	0.9634	0.8658	0.6057	0.2216

(Continued)

Table 10-3 (Continued)

	Numerator					*Denom.*	*P*						*P*			
Theta	*0*	*1*	*2*	*3*	*4*	*Sum*	*0*	*1*	*2*	*3*	*4*	*Total*	*1 or Better*	*2 or Better*	*3 or Better*	*4*
1.89	1.0000	2.6912	7.2427	10.804	6.2965	28.035	0.0357	0.0960	0.2583	0.3854	0.2246	1.0000	0.9643	0.8683	0.6100	0.2246
1.90	1.0000	2.7183	7.3891	11.134	6.5535	28.794	0.0347	0.0944	0.2566	0.3867	0.2276	1.0000	0.9653	0.8709	0.6143	0.2276
1.90	1.0000	2.7183	7.3891	11.134	6.5535	28.794	0.0347	0.0944	0.2566	0.3867	0.2276	1.0000	0.9653	0.8709	0.6143	0.2276
1.91	1.0000	2.7456	7.5383	11.473	6.8210	29.577	0.0338	0.0928	0.2549	0.3879	0.2306	1.0000	0.9662	0.8734	0.6185	0.2306
1.92	1.0000	2.7732	7.6906	11.822	7.0993	30.385	0.0329	0.0913	0.2531	0.3891	0.2336	1.0000	0.9671	0.8758	0.6227	0.2336
1.93	1.0000	2.8011	7.8460	12.182	7.3891	31.218	0.0320	0.0897	0.2513	0.3902	0.2367	1.0000	0.9680	0.8782	0.6269	0.2367
1.94	1.0000	2.8292	8.0045	12.553	7.6906	32.077	0.0312	0.0882	0.2495	0.3913	0.2397	1.0000	0.9688	0.8806	0.6311	0.2397
1.95	1.0000	2.8577	8.1662	12.935	8.0045	32.964	0.0303	0.0867	0.2477	0.3924	0.2428	1.0000	0.9697	0.8830	0.6352	0.2428
1.96	1.0000	2.8864	8.3311	13.329	8.3311	33.878	0.0295	0.0852	0.2459	0.3935	0.2459	1.0000	0.9705	0.8853	0.6394	0.2459
1.97	1.0000	2.9154	8.4994	13.735	8.6711	34.821	0.0287	0.0837	0.2441	0.3945	0.2490	1.0000	0.9713	0.8876	0.6435	0.2490
1.98	1.0000	2.9447	8.6711	14.154	9.0250	35.794	0.0279	0.0823	0.2422	0.3954	0.2521	1.0000	0.9721	0.8898	0.6476	0.2521
1.99	1.0000	2.9743	8.8463	14.585	9.3933	36.799	0.0272	0.0808	0.2404	0.3963	0.2553	1.0000	0.9728	0.8920	0.6516	0.2553
2.00	1.0000	3.0042	9.0250	15.029	9.7767	37.835	0.0264	0.0794	0.2385	0.3972	0.2584	1.0000	0.9736	0.8942	0.6556	0.2584
2.01	1.0000	3.0344	9.2073	15.487	10.175	38.904	0.0257	0.0780	0.2367	0.3981	0.2616	1.0000	0.9743	0.8963	0.6596	0.2616
2.02	1.0000	3.0649	9.3933	15.958	10.591	40.007	0.0250	0.0766	0.2348	0.3989	0.2647	1.0000	0.9750	0.8984	0.6636	0.2647
2.03	1.0000	3.0957	9.5831	16.444	11.023	41.146	0.0243	0.0752	0.2329	0.3997	0.2679	1.0000	0.9757	0.9005	**0.6676**	0.2679

NOTE: Step values are .9, .9, 1.49, and 2.43 for score points 1, 2, 3, and 4, respectively.

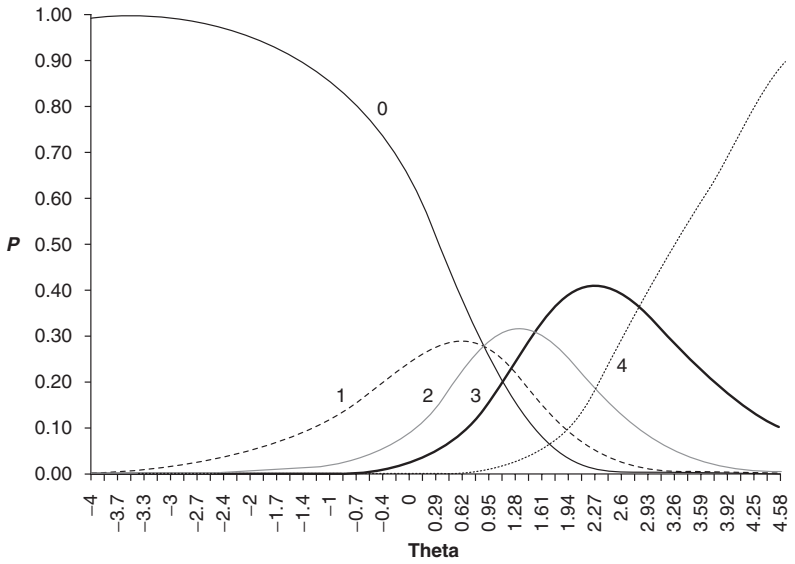

Figure 10-2 Response Characteristic Curves for Score Points 0–4 (based on data in Table 10-3, Rasch scaling)

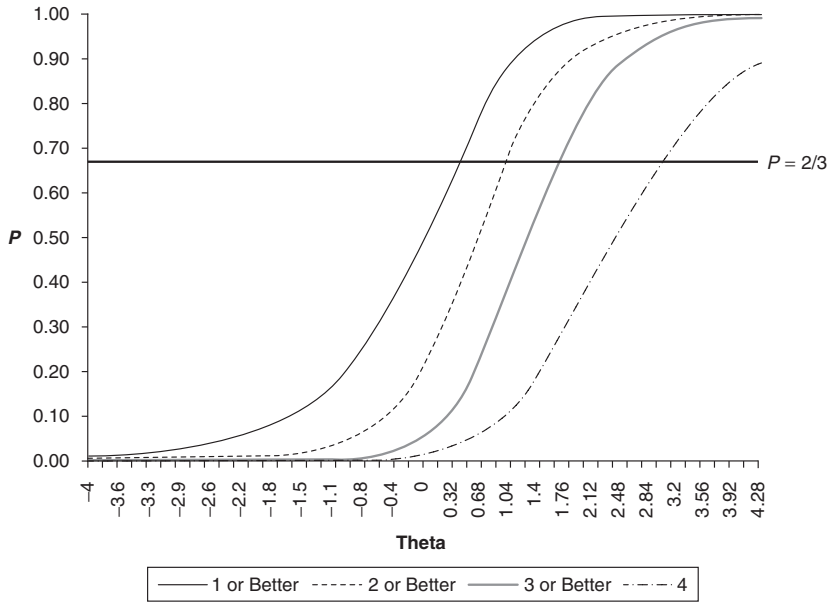

Figure 10-3 Probability of Obtaining a Given Score Point or Better as a Function of Ability (based on data in Table 10-3, Rasch scaling)

correctly. Thus, starting with a probability of 2/3 and solving for the ability (θ) needed to answer an item correctly, we obtain the following:

$$\theta = b_j + .693/1.7a_j \qquad \text{(Equation 10-24)}$$

(Again, had the RP been .67, rather than 2/3, the final result would have been $\theta = b_j + .708/1.7a_j$.)

For CR items, the situation becomes somewhat more complicated. Mitzel et al. (2001) used the two-parameter partial-credit (2PPC) model, with its fundamental equation relating the probability of obtaining score point k to student ability [$P_{jk}(\theta)$] and score point (step) difficulty (γ):

$$P_{jk}(\theta) = \exp(z_{jk})/\Sigma\exp(z_{ji}) \qquad \text{(Equation 10-25)}$$

where m_j is the number of score points or steps for item *j*,

$$z_{jk} = (k - 1)\alpha_j - \Sigma\,\gamma_{ji}; \qquad \text{(Equation 10-26)}$$

α_j is the discrimination index of item j;

k is the number of this score point or step; and

γ_{ji} is the step value for item *j* at step *i*.

Thus the probability of scoring at step *k* is a joint function of examinee ability, item discrimination, and the likelihood of obtaining any of the *k*-1 other scores. In this formulation, the value for a score of 0 (step 0) is set equal to zero; that is, $\gamma_{j0} = 0$ for all items. Procedures similar to those for establishing values of θ for each score point for each CR item within a Rasch framework can be established for the 2PL model.

As we illustrated in the Rasch context, we provide a portion of an Excel spreadsheet for a set of calculations, in this case for a hypothetical 3-point CR item, when a 2PL model is used. The spreadsheet appears as Table 10-4; the step values associated with each of the three score points are provided at the bottom of the table. And, also as before, we illustrate the response characteristic curves associated with each option for that item (Figure 10-4) and the curves associated with the probability of obtaining a given score or better on the item (Figure 10-5).

Directions to Bookmark Participants

As with other standard-setting methods, the selection and training of participants is an important aspect of the process. And, as with other

Table 10-4 Selected Spreadsheet Entries and Calculations for a Hypothetical 3-point CR Item, 2PL Scaling

	Numerator				*Denom.*	*P*				*P*			
Theta	*0*	*1*	*2*	*3*	*Sum*	*0*	*1*	*2*	*3*	*1 or Total*	*2 or Better*	*Better*	*3*
−4.00	1	0.087	0.002	0.000	1.089	0.918	0.080	0.002	0.000	1.000	0.082	0.002	0.000
−3.99	1	0.087	0.002	0.000	1.089	0.918	0.080	0.002	0.000	1.000	0.082	0.002	0.000
−3.98	1	0.088	0.002	0.000	1.090	0.917	0.081	0.002	0.000	1.000	0.083	0.002	0.000
−3.97	1	0.088	0.002	0.000	1.091	0.917	0.081	0.002	0.000	1.000	0.083	0.002	0.000
−3.96	1	0.089	0.002	0.000	1.091	0.916	0.082	0.002	0.000	1.000	0.084	0.002	0.000
0.14	1	1.371	0.549	0.074	2.993	0.334	0.458	0.183	0.025	1.000	0.666	0.208	0.025
0.15	1	1.380	0.556	0.075	3.011	0.332	0.458	0.185	0.025	1.000	**0.668**	0.210	0.025
0.16	1	1.389	0.563	0.077	3.030	0.330	0.459	0.186	0.025	1.000	0.670	0.211	0.025
0.17	1	1.399	0.571	0.078	3.048	0.328	0.459	0.187	0.026	1.000	0.672	0.213	0.026
2.15	1	5.239	8.013	4.115	18.367	0.054	0.285	0.436	0.224	1.000	0.946	0.660	0.224
2.16	1	5.274	8.120	4.198	18.593	0.054	0.284	0.437	0.226	1.000	0.946	0.663	0.226
2.17	1	5.310	8.229	4.283	18.822	0.053	0.282	0.437	0.228	1.000	0.947	0.665	0.228
2.18	1	5.345	8.340	4.370	19.055	0.052	0.281	0.438	0.229	1.000	0.948	**0.667**	0.229
2.19	1	5.381	8.452	4.458	19.291	0.052	0.279	0.438	0.231	1.000	0.948	0.669	0.231
2.20	1	5.417	8.565	4.548	19.531	0.051	0.277	0.439	0.233	1.000	0.949	0.671	0.233
4.39	1	23.342	159.047	363.925	547.313	0.002	0.043	0.291	0.665	1.000	0.998	0.956	0.665
4.40	1	23.498	161.183	371.280	556.961	0.002	0.042	0.289	0.667	1.000	0.998	0.956	**0.667**
4.41	1	23.655	163.348	378.784	566.787	0.002	0.042	0.288	0.668	1.000	0.998	0.957	0.668
4.42	1	23.813	165.541	386.440	576.795	0.002	0.041	0.287	0.670	1.000	0.998	0.957	0.670
4.43	1	23.973	167.764	394.251	586.988	0.002	0.041	0.286	0.672	1.000	0.998	0.957	0.672
4.44	1	24.133	170.017	402.219	597.370	0.002	0.040	0.285	0.673	1.000	0.998	0.958	0.673

NOTE: Step values are -0.33, 1.513, and 3.149 for score points 1, 2, and 3, respectively.

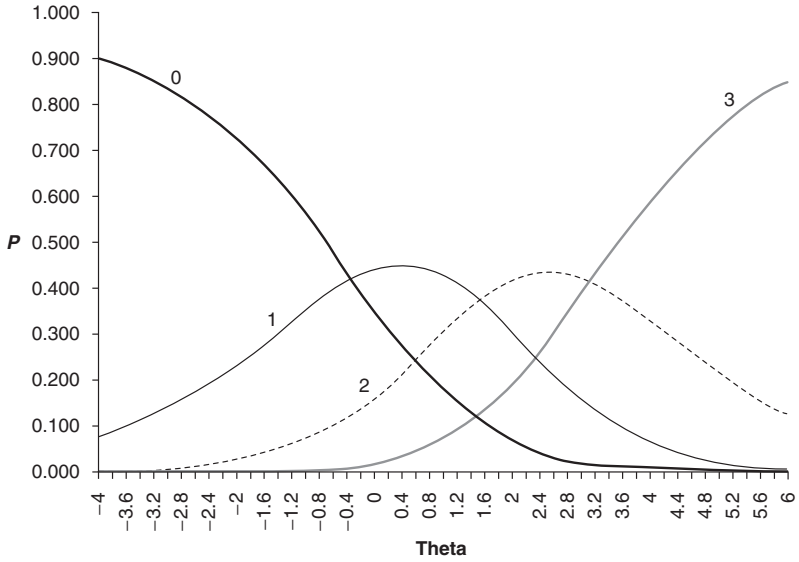

Figure 10-4 Response Characteristic Curves for Score Points 0–3 (based on data in Table 10-4, 2PL scaling)

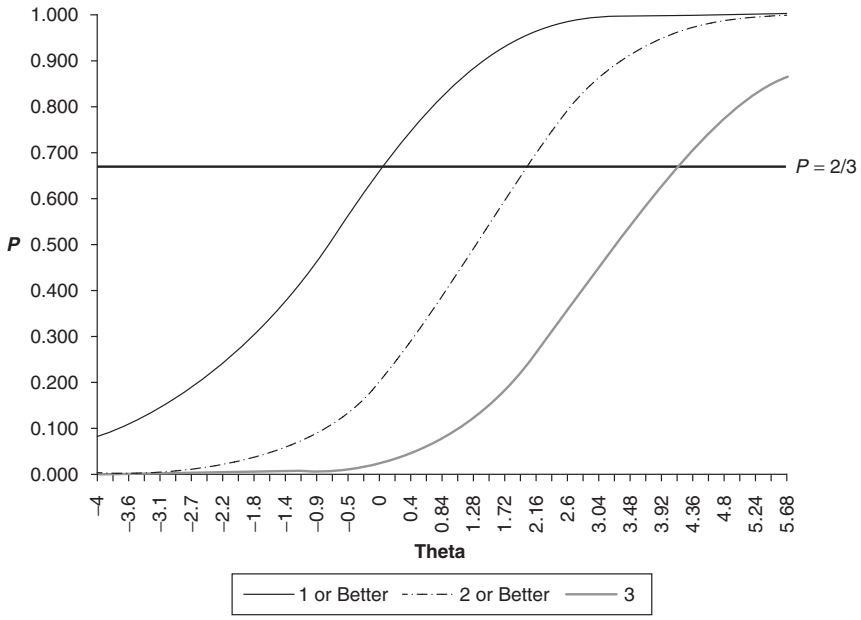

Figure 10-5 Probability of Obtaining a Given Score Point or Better as a Function of Ability (based on data in Table 10-4, 2PL scaling)

methods, when the Bookmark method is used participants must gain a clear understanding of the judgment task they are to perform.

The task presented to participants in a Bookmark standard-setting procedure is straightforward. Using the OIB assembled with one item (or score point) on each page, they are instructed to indicate the point at which they judge that the borderline or minimally qualified examinee's chances of answering the item correctly (or obtaining the score point) fall below the specified response probability or decision rule. For example, if a 2/3 decision rule is used, participants beginning to work through the OIB would ordinarily judge that the minimally qualified examinee would have better than a 2/3 likelihood of answering items at the beginning of the OIB (i.e., the easiest items) correctly. At some point in the OIB, however, participants would begin to discern that the chances of the minimally qualified examinee answering correctly approach and begin to drop below 2/3. Participants are instructed to indicate the point in the OIB at which the chances of the minimally qualified examinee answering correctly drop below 2/3. They indicate this judgment by placing a page marker—often a self-adhesive note or similar indicator—on the first page in the OIB at which the chance drops below the criterion. That is, the participants are indicating that the items prior to the marker represent content that the minimally qualified examinee would be expected to master at the RP or decision rule specified.

Standard-setting panelists generally work in small groups, evaluating the contents of small clusters of items as they appear in the difficulty-ordered test booklet. They discuss what makes one item or group of items more difficult than those that preceded it and ultimately place a bookmark at a point where they believe the difficulty of the subsequent items exceeds the ability of an identified group of students. In standard-setting contexts where more than one cut score is required (e.g., *Basic, Proficient,* and *Advanced*), participants would begin with the first item and ask themselves if a minimally qualified student (or group of students) at a particular achievement level (e.g., just barely *Basic*) would have the specified chance of answering the item correctly. They would then ask themselves the same question for each subsequent item until they reached one where they could not answer affirmatively. The final item yielding an affirmative response would mark the boundary of that performance level, and the participants would place a bookmark at that point (i.e., after the last attainable item). After making that judgment for the *Basic* category, participants would continue examining items beyond the bookmark just placed in order to identify the *Proficient* cut score, and so on for each cut score required.

Calculating Bookmark Cut Scores

Once participants have expressed their judgments by placing one (or more) bookmarks in the OIB, these judgments can be translated into cut scores. In a traditional Bookmark approach, the translation from bookmark placement to cut score is straightforward. For example, suppose that a participant has placed his or her bookmark on page 39 of a 50-page OIB to distinguish between *Proficient* and *Advanced* achievement levels. This does *not* correspond to a raw cut score of 39; rather, as we have indicated previously, this mark signals the participant's judgment that examinees classified as *Advanced* would be expected to be successful (defined in terms of whatever decision rule is being used) on the items through page 39 in the OIB. Of course, the examinee would also have some probability of success on the items after page 39. To obtain the cut score, the ability level associated with RP67 (or whatever decision rule is in place) that corresponds with the page in the OIB on which the bookmark was placed is the cut score, expressed in ability scale (i.e., theta) units. In this example, the theta associated with RP67 for the item appearing on page 39 of the OIB is the recommended cut score. From this point, it is a simple matter to transform the theta value to the raw score metric via the test characteristic curve or to another scaled score metric using the appropriate raw-to-scaled-score or theta-to-scaled-score conversion equation.

In the illustration in the preceding paragraph, the Bookmark cut score for a performance level was based only on the bookmark placement of a single participant. Obviously, in any application of the Bookmark standard-setting method, the procedure will be implemented using a large panel of participants. In the usual case, the bookmark placements of the participants will vary. In our experience, the typical method for addressing this situation is to find, for each participant, the theta (ability) level associated with the page in the OIB immediately preceding the one on which the participant's bookmark was placed in the same manner as just described. The result is a distribution of theta values, one for each participant. The overall recommended cut score in theta units is derived by taking the mean of these theta values and then obtaining the cut score in raw (or scaled score) units using one of the methods described in the preceding paragraph. We note, however, that the choice of central tendency measure most frequently used—that is, the mean—is perhaps based largely on statistical tradition and that the use of another statistic such as the median would likely be equally appropriate.

Finally, a point of clarification is in order here regarding the actual placement of the bookmark and the corresponding ability value that is used in cut score calculations. As we have described, when a Bookmark procedure

is conducted, participants are instructed to place their bookmark on the last page for which the participant could answer affirmatively the standard-setting question "Would an examinee just barely at this level have a 2/3 chance of answering this item correctly?" However, in other standard-setting sessions, participants are sometimes told to place their bookmarks on the first page for which the answer to this question is "No." If, for example, a participant answered "Yes" regarding the item on page 27 and "No" for the item appearing on page 28 of the OIB, some facilitators would have the participant place the bookmark on page 27, and some would have the participant place the bookmark on page 28. Strictly speaking, if the participant were actually placing a bookmark in a book, it would be placed between pages 27 and 28. In the example we have described here, the correct theta value for use in calculating the Bookmark cut score is the one found on page 27 of the OIB, not the one on page 28.

Our point here is that, regardless of the instructions given, it should be made clear to all involved (facilitators and participants) what is intended when a bookmark is placed in a given location: The most difficult item for which the participant can answer the standard-setting question affirmatively is the item whose values are entered into Bookmark cut-score calculations, whether participants identify that item by placing a bookmark on it, after it, or on the next page.

An Implementation of the Bookmark Procedure

Much of the terrain covered in previous sections of this chapter has outlined the mathematical foundations of the Bookmark standard-setting procedure. In this portion of the chapter, we seek to illustrate a typical Rasch-based application of the Bookmark method of the sort that is commonly used in the context of standards-referenced K–12 student achievement testing. In the illustration presented in the following paragraphs, we describe many practical aspects of the method, including training, presentation of the ordered booklet, and rounds of ratings.

Training

Training for a Bookmark standard-setting activity typically involves familiarizing participants with the performance level descriptions (PLDs), the test on which performance standards will be set, and the Bookmark standard-setting procedure. A sample agenda for a three-day session using the Bookmark method is shown in Figure 10-6. During the first day, participants

Day 1		
	8:00 A.M.	Registration, breakfast
	8:30	Introductions; distribute materials; collect security forms
	8:45	Background and overview
	10:00	Break
	10:15	Test administration
	12:30	Lunch
	1:30 P.M.	Test scoring and discussion
	3:00	Review of performance level descriptors
	4:00	Adjourn
Day 2		
	8:00 A.M.	Breakfast
	8:30	Distribute materials; introduction to the Bookmark procedure
	10:00	Break
	10:15	Practice round; evaluation of readiness
	11:00	Questions & answers
	Noon	Lunch
	1:00 P.M.	Instructions for Round 1
	1:15	Round 1
	3:45	Wrap-up
	4:00	Adjourn
Day 3		
	8:00 A.M.	Breakfast
	8:30	Distribute materials; review of Round 1 results
	9:45	Round 2
	Noon	Lunch
	1:00 P.M.	Discussion of Round 2 results
	1:30	Round 3
	3:00	Final recommendations
	3:30	Closure; evaluation
	4:00	Adjourn

Figure 10-6 Sample Agenda for Bookmark Standard-Setting Procedure

receive an overview of the purpose of the session and their objectives. This overview is followed by administration and scoring of the tests in order to give participants a clear understanding of the test contents. This activity is then followed by presentation and discussion of the PLDs. Placing the PLDs after the administration and scoring of the test helps participants view the content of the PLDs in a real-world context. Once participants understand and can articulate key components of the PLDs, they are given an opportunity to narrow the definition of each level to apply to those just barely in each performance level, that is, students at the threshold or cut score for that level.

On the morning of the second day, participants receive training in the specifics of the Bookmark method, followed by a short practice round in

which they place bookmarks for one cut score. A complete set of sample training materials that can be adapted to differing contexts is available at www.sagepub.com/cizek/bookmarktraining. After completing the practice activities, participants discuss their experiences and complete an evaluation form to assess their understanding of the training and their readiness to begin the bookmarking tasks.

Introducing the Ordered Item Booklet

As described previously, the OIB for a Bookmark standard-setting procedure consists of a series of SR and CR items in difficulty order, with the easiest item on the first page and the most difficult item on the last page. It is worth noting at this point that in the Rasch model it makes no difference whether the items are ordered by difficulty or by ability required to have a 2/3 chance of correct response; either method will result in the same ordering. If a 3PL or 2PPC model is used however, item difficulty and required ability will not necessarily order the items in the same way because the required ability is a function of both item difficulty and discrimination. Given two items of equal difficulty, the item with the lower discrimination index will require the higher ability to yield a 2/3 chance of correct response. (Recall that $\theta = b_j + .693/1.7a_j$, so that as a_j increases, the right side of the equation decreases.) Under these circumstances, a more difficult item might precede a less difficult item by several pages in an OIB ordered by theta, rather than by difficulty. In our experience, participants in a Bookmark procedure, who are usually far more sensitive to item difficulty than to discrimination, can be confused by an ordering based on theta values; thus it seems preferable to order booklets strictly by item difficulty.

Figure 10-7 shows an enlargement of a single page in an OIB. Information on the page includes page number, original item number and score point, and the Rasch achievement level required for a 2/3 chance to answer the item correctly. The key (A) is placed at the bottom of the page in a smaller font to serve as a quick check on the participant's own response to the item without interfering with the participant's estimation of the difficulty of the item. In practice, because items associated with a given stimulus (e.g., reading passages, graphics for sets of science or geography items, etc.) are likely to vary widely in difficulty and therefore be scattered throughout the test booklet, all common stimulus materials are placed in a companion booklet. The companion booklet is distributed to participants along with the difficulty-ordered test booklet.

The OIB page shown in Figure 10-7 contains all the information a participant would need to make a judgment about the item. All the information is printed at the top of the page so that it will be easily accessible to

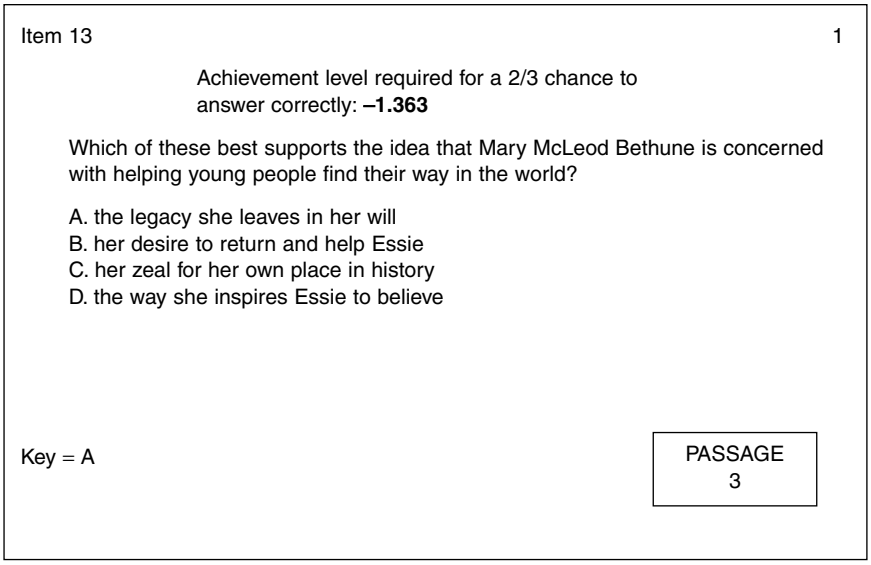

Figure 10-7 Sample Page From Ordered Item Booklet

participants. As can be seen in the figure, this item appears on page 1 of the OIB (as indicated by the numeral 1 in the box at the upper right corner of the page). This page number is boldfaced and of a larger size that makes it clearly distinguishable from other numbers on the page; this is important because participants use the page number as their indicator for a bookmark placement. The figure also shows that this item appeared as Item 13 in the actual test form, as indicated by the 13 printed in the upper left corner of the page. Also printed on the page is the achievement level (i.e., ability or theta) required for an examinee to have a 2/3 chance of answering this item correctly assuming (as is true in the sample page shown) that the item is an SR format item. In the sample page shown in the figure, the ability required (expressed in logits) is –1.363. If the item on this page had been a CR format item, the ability level expressed in logits would be the value associated with a 2/3 chance of obtaining that particular raw score point or higher. These values are obtained as described previously in this chapter.

Round One of a Bookmark Procedure

After an introduction to the procedure, each participant receives an OIB, a stimulus booklet, and a set of bookmarks. As mentioned earlier, the OIB

has one item per page, starting with the easiest item in the test booklet; each page contains information like that shown in Figure 10-7. Each CR item is represented once for each of its score points, as noted previously. Each CR page contains the item and one or more sample responses that are exemplars of the particular score point. Because there are several different ways to earn each score point, it is often a good idea to select sample responses that cover a broad range of possibilities across the various CR items.

For tests that have common stimuli (e.g., reading passages, maps, graphs), a separate stimulus booklet is prepared and distributed to participants. In an OIB, items for a given scenario, map, case, chart, or other stimulus are scattered throughout the booklet. To simplify the task participants face in matching items in the OIB with their associated stimuli, it is helpful to create a code for each stimulus and then repeat that code at the beginning of the corresponding item in the OIB. The box in the bottom right corner of Figure 10-7 provides a correspondence between that item and its associated stimulus (in this case, Passage 3).

Each participant also receives a printed form on which to enter his or her bookmarks (page numbers). The forms are printed on one side of a piece of card stock. Each form is similar to the one shown in Figure 10-8. In Rounds 1 and 2, participants enter the page number for each bookmark. At Round 3, participants will be familiar with the relationship between page number and cut score. At this stage, participants may enter page numbers and associated cut scores, as well as the impact data. The purpose of asking each participant to also enter the impact data is to help ensure that each participant is fully cognizant of the consequences that his or her recommendations will have in terms of the percentages of examinees that would be classified into each of the performance categories if the participants' cut scores were applied to actual test results.

During Round 1, participants usually work in small groups of three to five individuals. While they discuss the item contents among themselves, each participant completes his or her own Bookmark recording form like the one shown in Figure 10-8. As they complete Round 1, participants review their forms to make sure they are complete, return all materials to the facilitator, and are dismissed for the day.

Obtaining Preliminary Bookmark Cut Scores

At the end of Round 1 (and following rounds), standard-setting staff collect participants' bookmark cards and enter the values from the cards into a spreadsheet similar to the one shown in Table 10-5. After verifying the accuracy of the results, meeting facilitators return the cards to the

Panelist Number _____

Directions: Enter your Bookmark page numbers for each performance level in the spaces below.

ROUND 1

	Basic	***Proficient***	***Advanced***
Page Numbers			

ROUND 2

	Basic	***Proficient***	***Advanced***
Page Numbers			

ROUND 3

	Basic	***Proficient***	***Advanced***
Page Numbers			
Cut Scores			
% At or Above			

Notes:

Figure 10-8 Sample Bookmark Participant Recording Form

participants, along with the results. The sample information shown in Table 10-5 allows participants to see where their bookmarks fall relative to those of other participants. It also gives them a sense of where the group average lies, as well as how far their own bookmarks fall from the group average.

Table 10-5 provides a summary of bookmark placements, in addition to the resulting cut scores. Also shown are the mean cut score (along with its standard deviation), the minimum and maximum recommended cut scores for each performance level, and cut scores one standard deviation above and one standard deviation below the mean recommended cut scores. Individual cut scores in raw score units are not shown, but means, medians, minimum, and maximum cut scores in raw score units are provided.

Table 10-5 Sample Output From Round 1 of Bookmark Standard-Setting Procedure

	Basic		*Proficient*		*Advanced*	
Participant ID No.	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*
1	5	–0.334	12	0.286	46	1.627
2	8	0.082	22	0.600	46	1.627
3	8	0.082	22	0.600	47	1.650
4	6	–0.243	22	0.600	46	1.627
5	10	0.270	18	0.551	38	1.333
6	6	–0.243	16	0.493	39	1.340
7	9	0.193	21	0.579	40	1.489
8	6	–0.243	16	0.493	39	1.340
9	7	–0.176	16	0.493	40	1.489
10	8	0.082	16	0.493	40	1.489
11	7	–0.176	16	0.493	43	1.586
12	8	0.082	16	0.493	41	1.510
13	9	0.193	32	1.046	42	1.580
14	9	0.193	23	0.616	46	1.627
15	9	0.193	26	0.891	39	1.340
16	9	0.193	14	0.440	42	1.580
17	13	0.420	19	0.558	38	1.333
18	8	0.082	13	0.420	22	0.600
19	10	0.270	17	0.540	39	1.340
20	11	0.272	17	0.540	39	1.340
Summary Statistics in Theta (Ability) Metric						
Mean cut		0.060		0.561		1.442
Median cut		0.082		0.540		1.489
SD		0.217		0.161		0.233
Minimum		–0.334		0.286		0.600
Maximum		0.420		1.046		1.650
Mean – 1SD		–0.158		0.401		1.209
Mean + 1SD		0.277		0.722		1.676
Summary Statistics in Raw Score Metric						
Mean cut		22.04		28.73		39.87
Median cut		22.31		28.44		40.32
Minimum		18.00		25.00		30.00
Maximum		27.00		36.00		42.00
Mean – 1SD		19.46		26.51		37.34
Mean + 1SD		24.83		30.99		42.00

The translation of cut scores in the theta metric to a cut score in raw score units is a relatively straightforward process. Programs such as WINSTEPS or other IRT-based programs (e.g., PARDUX, PARSCALE, etc.) produce a conversion table showing raw scores and associated theta values. Using the calculated mean values for thetas at the three performance levels illustrated in Table 10-5, each of the three thetas is located in the conversion table and the closest raw score is obtained (or interpolated). Because a precise correspondence between the exact theta cut and raw cut will almost never be observed, the board or entity responsible for the performance standards, in advance of standard setting, will need to make a policy decision regarding whether to take the closest raw score, the raw score with an associated theta value just below the calculated mean theta, the raw score with associated theta value just above the calculated mean theta, or some other value. As we have urged previously, such decisions should also be documented, along with the rationale behind them.

A Caveat and Caution Concerning Bookmark Cut Scores

We digress for a moment from our description of this specific Bookmark implementation to offer a clarification and caution regarding how Bookmark cut scores are obtained. Indeed, we have seen a variety of applications of the Bookmark procedure in which alternative mechanisms for calculating a cut score have been employed. For example, in some applications of the Bookmark standard-setting procedure, the cut scores have been obtained by simply taking the mean recommended page number in the OIB and translating that number into a raw score. For instance, if the mean page number were 29, 29 could be taken as the cut score. The rationale for doing so would be that, on average, participants thought the minimally qualified examinee at that level would have a 2/3 chance of answering the first 29 items correctly. Such a procedure is ill-advised, however, and closer examination of the logic behind the appropriate procedure seems warranted.

The logic of setting a cut score at the raw score associated with the mean theta identified by participants is this: Participants place their bookmarks on the last item in the OIB for which they believe a minimally qualified examinee has a 2/3 chance of answering correctly. Minimally qualified examinees will still have some chance of answering subsequent items correctly, of course; right up to the end of the OIB, minimally qualified examinees (indeed nearly all examinees) will have some (very small) chance of answering each item correctly. Moreover, these examinees will have a greater than 2/3 chance of answering items correctly that appear prior to the location of their bookmarks.

IRT models are based on the notion that each examinee has a calculable probability of answering each item correctly (or obtaining any given score point on a CR item). The estimated raw score for a given theta is the sum of these probabilities and expected values. Thus, for example, if the mean theta value for *Proficient,* based on the estimates of the 20 participants represented in Table 10-5, is .561, then the Bookmark-based cut score for *Proficient* is 28.73, which is the interpolated value from the WINSTEPS output showing theta values associated with raw scores of 28 and 29. If we simply took the average page number as our cut, we would get a cut score of 18.7. If this value were used as the cut score (rounded to either 18 or 19; for purposes of this point, it doesn't matter which), there would be approximately a 10-raw-score point difference between this value and the correct value of 28.73—clearly a practically significant difference.

Round One Feedback to Participants

An example of one type of normative information provided to participants in a Bookmark standard-setting procedure is shown in Figure 10-9. At this point, only the page numbers bookmarked by the participants are shown. In this way, participants get a graphic view of how their bookmarks compare to the bookmarks of the other participants. The figure helps illustrate where there are gaps, that is, page ranges in which no participant chose to place a bookmark for any cut score. In subsequent rounds, the page ranges will typically not be the focus of attention; rather, discussion and consideration will center on the range of pages in which Round 1 judgments have indicated that the eventual cut score recommendations will likely be located.

Interestingly, Figure 10-9 also shows where there are overlaps in individual judgments. For example, one participant placed his bookmark at page 37 for the *Advanced* level, whereas another participant placed her bookmark for *Proficient* on page 39. In effect, one participant would set the cutoff for *Proficient* higher than at least one participant would set the cutoff for *Advanced*. Visualizations such as Figure 10-9 are excellent mechanisms for promoting the important discussions that will characterize Rounds 2 and 3.

In addition to normative information, impact information is also usually provided to participants in any standard-setting procedure. The juncture at which such information is provided varies, however. In this case, we illustrate the provision of impact information at the end of Round 1, although it can be introduced at the end of Rounds 1, 2, or 3. We note, however, that in our experience the later that impact information is presented to participants, the less an impact on participants' judgments it appears to have. The purpose of impact data is to allow participants to see how many (or what

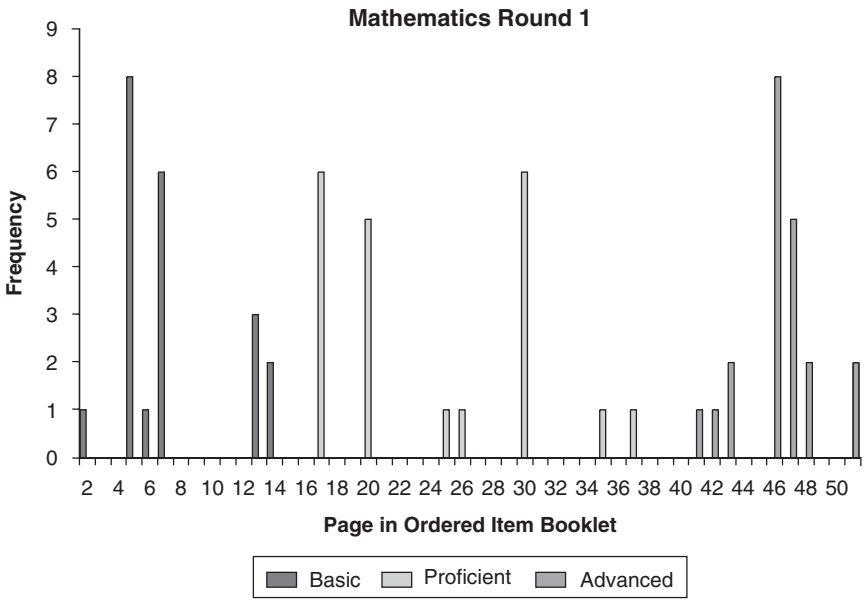

Figure 10-9 Sample Display of Round One Bookmark Placement Feedback

percentage of) examinees would be classified at each performance level if the mean cut scores from that round were implemented. An example of impact information is presented later in this chapter (see Table 10-6).

Round Two of a Bookmark Procedure

Following the schedule shown previously in Figure 10-6, the third day of the standard-setting session begins with participants receiving their OIBs and other materials from Round 1 plus the bookmark summary data and impact information from Round 1. The first activity is a discussion, led by meeting facilitators and centered on the Round 1 ratings and impact data. This discussion generally focuses on range of cut scores, areas of particular disagreement, and concerns about difficulty location of individual items. As part of this discussion, it is sometimes helpful for participants to explicitly address differences between their perceived difficulty of a particular item and the placement of that item relative to others in the OIB.

Once participants have discussed the results of Round 1 as a total group, they continue their work in small groups of three to five members to begin Round 2. The reassignment of participants to smaller groups may be random, or it may be done purposefully in order to bring divergent points

of view together at the same table. In either event, reassignment between rounds maximizes opportunities for participants to express their own—and hear others'—points of view. The participants' task for Round 2 is essentially identical to that of Round 1, consisting of (re)consideration of bookmark placements and the content of items captured by performance levels and discussion of those judgments with other small group members. The primary difference between Rounds 1 and 2 is the amount of information available to each participant. At the end of the second round of ratings, facilitators collect all materials and dismiss participants for lunch, during which facilitators again analyze the bookmark placements and prepare reports similar to those shown in Table 10-5 and Figure 10-9. This information is provided to participants at the beginning of Round 3.

Round Three of a Bookmark Procedure

To begin Round 3 of a Bookmark standard-setting procedure, participants again use their OIBs and are provided with all of their other Round 2 materials plus a summary of the Round 2 judgments. In our experience, it is at this point that a special version of Table 10-5 appears to be quite helpful to participants. An example of this version is shown in Table 10-6. The distinctive feature of Table 10-6 is that it includes actual raw score equivalents associated with the theta values that are the recommended cut scores. This feature helps clarify for participants the relationship between their bookmark placements, the theta values associated with those placements, and the impact that a bookmark placement (or changing a bookmark placement) will have on both the raw cut score and the percentages of examinees classified at or above a given performance level.

Round 3 begins with facilitators' leading a discussion of the impact data and other topics of concern from Round 2. At the end of this discussion, participants are asked to evaluate all of their previous ratings and all information at hand and to simply enter three bookmarks and the associated cut scores on their recording form (see Figure 10-8). At this stage, participants are actually asked to enter several pieces of data on their recording forms. Reviewing Figure 10-8 reveals that, in addition to the page number at which they have placed bookmarks for each performance level, participants are asked to enter the raw cut score associated with the page number and the corresponding percentage of examinees that would be classified at or above that level. The requirement that participants enter all three of these values for each cut score is an attempt to verify participants' understanding of the final task, to highlight the impact of the judgments, and to provide a check on the accuracy of the participants' intentions. The final task of

Table 10-6 Round 3 Feedback for Bookmark Standard-Setting Procedure

Page No. in OIB	*Original Item No.*	*IRT Item/Step Difficulty*	*Theta @ RP*	*Raw Cut Score*	*% At or Above*
1	6	–2.305	–1.612	8	99.61
2	2	–1.986	–1.293	10	99.11
3	3	–1.950	–1.257	10	99.11
4	12	–1.304	–0.611	15	95.66
5	14	–1.027	–0.334	18	92.13
6	11	–0.936	–0.243	19	90.62
7	1	–0.869	–0.176	20	88.86
8	28	–0.611	0.082	23	82.34
9	24	–0.500	0.193	24	79.66
10	15-1	0.480	0.270	25	76.47
11	18	–0.421	0.272	25	76.47
12	17	–0.407	0.286	25	76.47
13	26-1	0.790	0.420	27	68.80
14	5-1	0.650	0.440	27	68.80
15	8-1	0.350	0.440	27	68.80
16	7	–0.200	0.493	28	64.77
17	29-1	0.240	0.540	29	60.69
18	27	–0.142	0.551	29	60.69
19	34	–0.135	0.558	29	60.69
20	20	–0.124	0.569	29	60.69
21	13	–0.114	0.579	29	60.69
22	37-1	1.250	0.600	30	56.48
23	21	–0.077	0.616	30	56.48
24	15-2	0.750	0.740	32	47.81
25	26-2	0.320	0.810	33	43.56
26	16	0.198	0.891	34	39.34
27	33-1	0.620	0.900	34	39.34
28	37-2	0.090	0.910	34	39.34
29	30	0.252	0.945	35	35.18
30	36	0.295	0.988	35	35.18
31	38	0.305	0.998	35	35.18
32	4	0.353	1.046	36	31.15
33	15-3	0.320	1.090	36	31.15
34	35	0.464	1.157	37	27.20
35	9	0.498	1.191	38	23.39
36	5-2	0.050	1.200	38	23.39
37	26-3	1.650	1.290	39	19.64
38	10	0.640	1.333	39	19.64
39	32	0.647	1.340	39	19.64
40	19	0.796	1.489	41	12.45
41	37-3	1.530	1.510	41	12.45
42	8-2	0.630	1.580	42	9.61
43	22	0.893	1.586	42	9.61
44	31	0.896	1.589	42	9.61
45	26-4	–0.010	1.590	42	9.61
46	23	0.934	1.627	42	9.61
47	25	0.957	1.650	42	9.61
48	15-4	1.130	2.040	45	3.73
49	37-4	0.860	2.080	45	3.73
50	29-2	1.280	2.120	46	2.69
51	33-2	1.560	2.410	47	1.71

Round 3 occurs as participants complete evaluation forms (see Chapter 3) and are dismissed. Facilitators then check each completed bookmark for accuracy, tally these final ratings, and calculate the mean recommended cut score for each achievement level.

Alternative Procedures and Limitations

Given the time that has elapsed since its introduction, the limited amount of published research evidence about the Bookmark procedure is somewhat surprising (see Karantonis & Sireci, 2006). Then again, given the number of times the Bookmark procedure has been used, it is not surprising that, for the most part, many initial concerns about the method have been addressed via procedural changes based on these experiences. Yet one fundamental aspect of the procedure must be reckoned with each time it is employed: The cut score is absolutely bound by the relative difficulty of the test. It is this limitation that future research on the Bookmark method must address.

To see the impact of this limitation, we can consider the relative alignment of the difficulty of a test and the ability of the population of examinees who will take the test. If the test is easy, relative to the examinee population, it will be impossible to set a cut score below a certain point, no matter what any of the participants may wish. For example, let us assume that every participant placed a bookmark for Basic on page 1 of an OIB. On this (relatively) easy test, the ability level associated with an RP67 will very likely yield a cut score of two (or more) on the raw score scale. In a recent application, we found that the theta level for page 1 in a certain OIB yielded a raw score of 10! Similarly, a bookmark placed even on the last page of an OIB will not necessarily yield a raw cut score of 100% correct. We have even conducted bookmark procedures in which some participants would have preferred to go beyond the last page in the booklet for their final bookmark, claiming that the most difficult item in the booklet was not sufficiently difficult to distinguish the highest category of performance.

Of course, tests that are too easy or too difficult for the population of examinees present problems that are not unique to the Bookmark standard-setting method, but would pose problems for other methods as well. Indeed, this limitation may actually be a credit to the method in that it brings the limitation to light. To address the limitation, careful item writing and test construction procedures must be in place; standard-setting methods cannot compensate for weaknesses in content coverage, performance characteristics of items, and so on.

A second difficulty that arises in Bookmark applications has to do with unusually large gaps in difficulty between items. When the OIB comprises

items selected from a deep item pool (as opposed to a specific test form), this problem can generally be avoided. However, when the ordered item booklet is created directly from the operational test, it is likely that the dispersion of item difficulties will be uneven. This problem may cause difficulty for participants in the process, particularly when it is apparent that one of the cut scores falls in one of the gaps. For example, let us assume that five contiguous items in the OIB have the following RP67 theta values: Item 21 (1.41), Item 22 (1.53), Item 23 (1.62), Item 24 (2.04), and Item 25 (2.17). Participants examining this string of items may judge that Item 23 is well within the grasp of a barely *Proficient* examinee, but that Item 24 is far beyond the grasp of such an examinee. In such a case, participants may not wish to place their bookmark on either item, preferring instead—if it were possible—to place a bookmark somewhere between Items 23 and 24. Their predicament is to settle for placing a bookmark on Item 23, which may yield a cut score that is lower than the panel as a whole can support, or placing the bookmark on Item 24, which could result in a cut score higher than participants are comfortable recommending.

For these reasons, we recommend that special care be given to the development of the operational test if it will be used for Bookmark standard setting and that standard-setting issues be carefully considered early in the test-development process. It may be possible to forestall both problems (difficulty/ability mismatch and item difficulty gaps) through targeted test design. We recognize, of course, that final values for examinee ability and item characteristics can be known only after operational administration, making it especially critical that practitioners be aware of this potential problem and plan to avoid its consequences in advance.

Alternative procedures can be conceived to address these potential limitations, however. With regard to the item difficulty/examinee ability mismatch leading to unanticipated cut score placements, one simple approach essentially ignores the *b*-theta relationship. If this alternative is used, the page number in the OIB is taken directly as the raw cut score; that is, if a participant puts a bookmark on page 10, the recommended cut score is 10 points. Precisely this strategy was implemented in a study by Buckendahl, Smith, Impara, and Plake (2002); the authors reported that it worked well in the context of setting standards for a seventh-grade mathematics assessment in a midwestern school district.

Earlier in this chapter we described such an approach as an incorrect implementation of the Bookmark method. However, this alternative (a "modified Bookmark" method?) does provide a simple strategy and a reasonable alternative provided that all related training, materials, feedback, and so on are similarly realigned. The study cited in the previous

paragraph provides only limited support for this alternative; however, more research would be required before its use can be recommended.

Another alternative involves using classical items statistics (i.e., *p* values) instead of IRT values to order the OIB. Then, for each page in the OIB, a scale score can be assigned to each page in the OIB such that, for example, page 10 would have a scale score equivalent to 10 raw score points. Although this strategy appears to remedy the difficulty-ability mismatch, it also raises new questions. In essence, this approach resets the scale scores in ways that can have unforeseen consequences, and before recommending this strategy we await the results of research that will uncover the intended and unintended consequences of this ordering strategy. Another alternative—and one that research is needed to address—also involves the ordering of the OIB. In all applications of the Bookmark method we are aware of, the items in the OIB are compiled in increasing difficulty order. Sequencing items in the opposite order (i.e., from hardest to easiest) seems like a plausible alternative; research evidence that either ordering produces similar cut scores would add validity support for the method.

With regard to the item difficulty gap problem, as we indicated previously, a specially constructed OIB created from a bank with an abundance of items at every difficulty level is technically preferable. However, as we also mentioned, grounding standard setting in an actual operational test is also highly desirable. Between these options, there may be a midpoint. Should the operational test yield gaps that are likely to interfere with setting standards via the Bookmark procedure, it would seem prudent to identify the location of those gaps and insert a small number of items from the bank to supplement the operational test form and ameliorate the gaps. The key consideration here to be weighed is the tradeoff between measurement precision and fidelity to the operational form. Particularly if the OIB only added (a small number of) items to and took no items away from the operational booklet, the objection to this practice might be easily overcome. As with many of the other decision points we have illustrated, this is a policy issue that would need to be addressed early in the planning of the standard setting.