The Alberta Journal of Educational Research *Vol. 52, No. 1, Spring 2006, 36-52*

Jie Lin
University of Alberta

The Bookmark Procedure for Setting Cut-Scores and Finalizing Performance Standards: Strengths and Weaknesses

The Bookmark standard-setting procedure was developed to address the perceived problems with the most popular method for setting cut-scores: the Angoff procedure (Angoff, 1971). The purposes of this article are to review the Bookmark procedure and evaluate it in terms of Berk's (1986) criteria for evaluating cut-score setting methods. The strengths and weaknesses of the Bookmark are critically examined and discussed. In general, the strengths of the Bookmark method are that it (a) accommodates constructed-response as well as selected-response test items; (b) efficiently accommodates multiple cut-scores and multiple test forms; and (c) reduces cognitive complexity for panelists. Despite unresolved issues like the choice and understanding of the response probability, the Bookmark method remains a promising procedure for setting cut-scores and finalizing performance standards.

La procédure de normalisation Bookmark a été développée pour aborder ce que l'on percevait comme étant les problèmes relatifs à la méthode la plus communément employée dans l'établissement des notes de passage : la procédure Angoff (Angoff, 1971). Les objectifs de cet article sont d'examiner la procédure Bookmark et de l'évaluer selon les critères établis par Berk (1986) pour évaluer les méthodes d'établissement de notes de passage. L'on discute, de façon éclairée, les forces et les faiblesses de la procédure. De façon générale, les forces de la méthode Bookmark sont les suivantes : (a) elle s'adapte aux questions à réponse construite aussi bien qu'aux questions à réponse choisie ; (b) elle admet efficacement plusieurs notes de passage et diverses formes d'examens ; et (c) elle amoindrit la complexité cognitive pour les juges. Malgré les questions non résolues telles le choix et la compréhension de la probabilité des réponses, la méthode Bookmark reste une procédure prometteuse dans l'établissement des notes de passage et la mise au point des normes de rendement.

Setting standards is becoming increasingly important due to the reform of standard-based education and growing public demands for educational accountability. Among the various methods of setting cut-scores corresponding to specified performance levels, the Angoff procedure (Angoff, 1971) is often considered "the industry standard" (Zieky, 2001). However, dissatisfaction with this procedure has been mounting. First, the Angoff procedure was designed for multiple-choice item formats and does not accommodate constructed-response item types very well. Second, the Angoff procedure involves the estimation of the percentage of the population who would respond correctly to each item for all the items and every performance standard level. The estimation of numerous *p* values can be tedious (Mitzel, Lewis, Patz, & Green, 2001). Poor accuracy has also been found in the estimation of item difficulty.

Jie Lin is a doctoral student in the Centre for Research in Applied Measurement and Evaluation, Faculty of Education. Her areas of interest include test adaptation, differential item functioning, standard setting, test equating, and second-language assessment.

Typically, panelists tend to overestimate performance on difficult items and underestimate performance on easy items (Bejar, 1983). Last, it is still questionable whether panelists are really capable of performing the Angoff tasks. Shepard, Glaser, and Bohrnstedt (1993) argue that the Angoff method is "fundamentally flawed" because the cognitive task of estimating the probability that a hypothetical student at the boundary of a given achievement level will get a particular item correct is nearly impossible. The panelists may instead be "doing the much simpler task of expressing their own opinion about how well an examinee would have to do to be considered minimally acceptable" (Zieky, 2001, p. 36). For a more thorough critical review of the Angoff and modified Angoff procedures, see Ricker (in press, this issue).

To address the perceived problems of the Angoff procedure, Lewis, Mitzel, and Green (1996) developed the Bookmark procedure for setting cut-scores and refining performance standards. The Bookmark procedure aims to (a) simultaneously accommodate selected-response and constructed-response test formats, (b) simplify the cognitive complexity required of panelists participating in standard-setting, (c) connect the judgment task of setting cut-scores to the measurement model, and (d) connect test content with performance level descriptors (Mitzel et al., 2001). Since its introduction in 1996, 31 states in the United States have implemented the Bookmark procedure to set cut-scores on large-scale assessments (Wisconsin Department of Public Instruction, 2003). So far, the publications and conference presentations on the Bookmark procedure have been mostly produced by the developers and their colleagues (Lewis, Green, Mitzel, Baum, & Patz, 1998; Lewis, Mitzel, Green, & Patz, 1999; Mitzel et al., 2001). The intent of this article is to provide an independent review of the Bookmark procedure, evaluate it in terms of Berk's (1986) criteria for evaluating methods for setting cut-scores, and critically examine the strengths and weaknesses of the procedure.

Basic Assumptions of the Bookmark Procedure

The Bookmark procedure is based on item response theory (IRT, Lord, 1980), a framework that characterizes the proficiency of examinees and the difficulty of test items simultaneously. Each IRT-scaled item can be represented by an item characteristic curve (ICC) that displays the relationship between the proficiency of an examinee and the probability of a correct response on an item (see Figure 1). IRT makes it possible to order items by the ability or scale score needed to have a specific probability of success. Items are thus mapped to locations on the IRT scale such that students with scale scores near the location of specific items can be inferred to hold the knowledge, skills, and abilities required to respond successfully to those items with the specified probability.

For the Bookmark procedure, the specified probability of success is set to 0.67; students with a scale score at the cut-point will have a 0.67 probability of answering the item at that cut-score correctly. The use of 0.67 as the response probability (RP) has been supported by the research of Huynh (1998). Huynh suggested that "the information function should be maximized based on (correct) response probabilities, because this is where examinees can be expected to have the requisite skills underlying a correct response" (Mitzel et al., 2001, p. 262). For the three-parameter logistic (3PL) IRT model (Lord, 1980; Lord & Novick, 1968), the item information function is maximized at θ for which

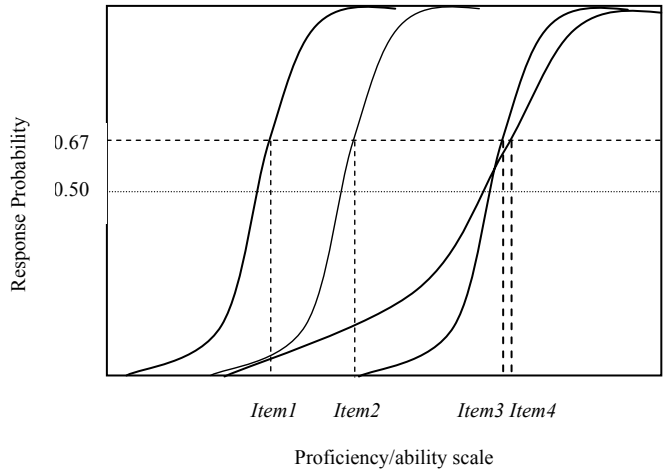

Figure 1. Item characteristic curves (ICCs) for selected-response items mapped at RP = 0.67. Adapted from Mitzel et al. (2001), p. 261.

$P(\theta)=(c+2)/3$ (Huynh, 1998). When guessing is removed ($c=0$) as a possible noise factor from the panelists' evaluation process, the value of RP equals 2/3. As Mitzel et al. (2001) put it, "the Bookmark task is based on a mastery judgment of what a student should know, not what a student might attain through guessing" (p. 261).

The three-parameter logistic model and the two-parameter partial credit model (Muraki, 1992) are used to calibrate the item parameters for selected-response items and constructed-response items respectively. Then the selected-response and constructed-response items are scaled jointly using computer programs such as MULTILOG (Thissen, 1999), PARDUX (Burket, 1991), or PARSCALE (Muraki & Bock, 1991). This joint scaling enables panelists to consider all the items together, whether selected-response or constructed-response, and to set a single cut-score for every performance level.

Overview of the Bookmark Procedure

The Bookmark Materials

In addition to commonly used materials like operational test booklets, student exemplar papers, and the scoring guide, the ordered-item booklet and its associated item map are central to the Bookmark procedure. Using the difficulty index (b parameter), the items are ordered from the easiest to most difficult in an item booklet. As illustrated in Figure 2, the ordered-item booklet has one item per page, with the first item being the easiest and the last item the hardest. The prompts for the constructed-response items appear multiple times throughout the ordered-item booklet, once for each score point. Similar to selected-response items, the location of a given constructed-response score point is defined as the point on the ability scale for which students have a 0.67 probability of achieving that item score or above. By scaling selected-response and constructed-response item score points together, both item types are placed into a single ordered-item booklet and thus are considered jointly by panelists (Mitzel et al., 2001). The purpose of the ordered-item booklets, as stated by Lewis et al. (1998), is "to help participants foster an integrated

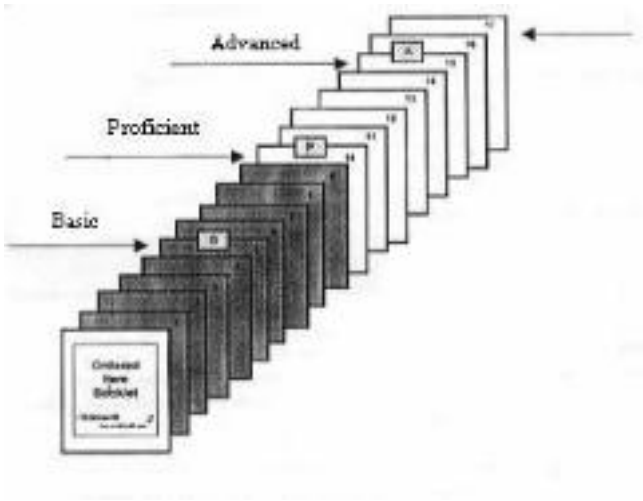

Figure 2. Illustration of ordered item booklet for the Bookmark procedure. Adapted from Mitzel et al. (2001), p. 256.

conceptualization of what the test measures, as well as to serve as a vehicle to make cut-score judgments" (p. 7).

Along with the ordered-item booklets, item mapping rating forms are provided as a guide to the booklets. The rating forms list all the items in the order in which they appear in the ordered-item booklets, with associated information such as the items' scale location, item number in the operational test, the standard or objective to which the item is referenced, space for the panelists to record their thoughts about what each item measures and why it is harder than the preceding items, and space for the panelists to enter their estimated cut-scores for each round (Lewis et al., 1998).

Panel Composition and Training

Each panel should consist of at least 18 panelists for each grade/content area (Lewis et al., 1998), although 24 panelists are recommended (Lewis et al., 1999). The panelists should be "representative of the state (or district) in terms of geographic location, socioeconomic status, ethnicity, gender, and community type (urban, suburban, rural)" (Lewis et al., 1999, p. 25). Typically the full panel is divided into three or four small groups so as to allow for greater discussion among panelists. Each small group consists of five to seven members.

Small-group leaders. The training of the group leaders (each leading a small group) involves a review of the standard-setting schedule and specific leadership responsibilities such as facilitating group discussion, keeping the group focused on the task, and watching the time for the group (Lewis et al., 1999).

Panelists. During the training session for panelists, a brief review is provided on "(a) the background and purpose of the testing program, (b) the content standards, (c) the general and/or specific performance-level descriptors, and (d) the stakes associated with the assessment and the performance levels (for students, teachers, schools, and districts)" (Lewis et al., 1999, p. 31). The importance of setting the cut-scores is also emphasized to panelists. Working as a large group, the panelists then take the test, examine selected-response

items and each score point of the constructed-response items, and review the scoring rubrics. A typical Bookmark conference agenda is provided in Table 1.

Setting bookmarks typically involves three rounds or iterations. Each round is intended to help increase consensus and reduce differences among the panelists.

Round 1. The main goals for Round 1 are to get panelists familiar with the ordered-item booklet, set initial bookmarks, and then discuss the placements. In this round panelists, working in their small groups, discuss what each item measures and what makes an item more difficult than the preceding item. The general performance descriptors for different levels (e.g., basic, proficient, and advanced) are also presented and discussed. Panelists are subsequently asked to discuss and determine the content that students should master for placement into a given performance level. Their independent judgments of cut-scores are expressed by simply placing a bookmark between the items judged to represent a cut-point. One bookmark is placed for each of the required cut-points. Items preceding the participant's bookmark reflect content that all students at the given performance level are expected to know and be able to perform successfully with a probability of at least 0.67. Conversely, these students are expected to perform successfully on the items behind the bookmark with a probability less than 0.67.

Table 1
Typical Bookmark Conference Agenda

Day	*Activity*
1	AM: Train large panel leaders PM: Train group leaders
2	AM: (in large group) Take test Review selected-response items Review constructed-response items in each score point PM: (in small groups) Review ordered item booklets Round 1 bookmark placement
3	AM: (in small groups) Present round 1 judgments Discuss bookmark placement within group Round 2 bookmark placement PM: (in large group) Present round 2 judgments with impact data Discuss bookmark placement Round 3 bookmark placement Present round 3 result with impact data (optional) Complete evaluation forms
4	AM: First and second drafts of descriptor writing PM: Final draft of descriptor writing

Note: Adapted from Mitzel et al. (2001), p. 253.

Round 2. The first activity in Round 2 involves having each member place bookmarks in his or her ordered-item booklet where each of the other panelists in their small group made their bookmark placement. For a group of six people, each panelist's ordered booklet will have six bookmarks for each cut-point. Discussions focus on the items between the first and last bookmarks for each cut-point. On completion of this discussion, the panelists independently reset their bookmarks. The median of the Round 2 bookmarks for each cut-point is taken as that group's recommendation for that cut-point.

Round 3. Round 3 typically begins with the presentation of impact data to the large group. The percentage of students falling into each performance level is presented given each group's median cut-score from Round 2. With this information of how students actually performed, the panelists discuss the bookmarks in the large group and then independently make their Round 3 judgments of where to place the bookmarks. The median for the large group is considered the final cut-point for a given performance level.

Definition of Cut-Scores

As mentioned above, the ability scores at the response probability of 0.67 obtained using IRT models are on a scale with a mean of zero and standard deviation of one. To avoid negative values for student scale scores and eliminate the need for decimal points in reporting student achievement, the theta scores are typically transformed onto a scale with a mean of 500 and standard deviation of 100. The scale location of the item immediately before the final cut-point is used as the operational cut-score for that particular level.

Finalizing Performance Standards

Based on the final cut-scores set, performance level descriptors are written by selected panelists. Performance descriptors describe the specific knowledge, skills, and abilities held by students at a given performance level. Generally, items before the bookmark(s) reflect the content that students at this performance level are expected to be able to answer correctly with at least a 0.67 likelihood. The knowledge and skills required to respond successfully to these items are synthesized to formulate the descriptors for this performance level. Performance level descriptors thus become a natural extension of the cut-score setting procedure.

Evaluation of the Bookmark Procedure Using Berk's Criteria

Berk's (1986) criteria for defensibility of standard-setting methods include two types of criteria: technical and practicable. Technical adequacy refers to "the extent to which a method satisfies certain psychometric and statistical standards that would render it defensible to experts on standard setting" (p. 140). However, a technically defensible procedure does not necessarily warrant a feasible procedure in that the procedure may be hard to implement or interpret. Therefore, the practicability of a procedure must also be taken into account. According to Berk, practicability refers to "the ease with which a standard-setting method can be implemented, computed, and interpreted" (pp. 143-144). The Bookmark procedure was evaluated against Berk's criteria using a three-point Likert scale, with 1 = not met, 2 = partially met, and 3 = fully met.

Technical Adequacy

1. The method should yield appropriate classification information (Rating: 3). The cut-scores produced by the Bookmark method permit dichotomous classification decisions at each cut-score. For example, students with scale scores higher than the cut-score for the proficient level are considered *proficient*, whereas those who score lower than the cut-score are classified as *non-proficient*.

2. The method should be sensitive to examinee performance (Rating: 3). The Bookmark method is sensitive to examinee performance in that it combines the content review process with the actual examinees' test results. The ordered-item booklet, a core component of the Bookmark approach, is based on the difficulty parameters estimated from the performance of the examinee population. That is to say, examinee performance somewhat determines the order of the items in the booklet, which then plays a crucial role in setting the final cut-scores.

3. The method should be sensitive to the instruction or training (Rating: 3). The Bookmark method is sensitive to the instruction or training received by examinees. If students were not taught the skills necessary to answer some of the items correctly, they would be more likely to perform poorly on these items. That is, the difficulty parameters of these items would tend to be relatively high at the response probability of 0.67. Therefore, in the ordered-item booklet, these items would probably be positioned more toward the back rather than the front of the booklet. The bookmarks would be more likely to be set before them, meaning that these uncovered items might not be included in the performance level descriptors and therefore not be considered something the students needed to master to achieve a given standard. Similarly, items addressing the content covered well in the classroom would be more likely to be answered successfully by the majority, achieve lower difficulty values, and therefore be classified as content that a student has to master to achieve a certain level. In this way the instruction or training received by examinees is somewhat reflected in the order of the items, which may then influence the setting of cut-scores.

4. The method should be statistically sound (Rating: 3). The Bookmark method is statistically sound in terms of the use of IRT models and the calculation of cut-scores and standard errors. When there is a close data-model fit, IRT models provide reliable estimates of item difficulty parameters, which form the base of the Bookmark procedure. In accordance with Standard 4.19 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), a measure of the variability among the panelists' judgments is provided in the bookmark procedure. Specifically, the cluster sample standard error (Cochran, 1963) is typically calculated from the Round 2 small-group medians. Because the panelists are divided into roughly balanced groups that work independently from Round 1 to Round 2, this cluster sample standard error reflects the stability of consensus in Bookmark cut-scores across independent small group replications (Lewis et al., 1998). Therefore, it seems fair to say that the Bookmark procedure is statistically sound in terms of the use of IRT models and the calculation of cut-scores and standard errors. The only concern, however, would be the selection of response

probability, that is, to what extent the cut-score may be manipulated by changing the response probability value. This concern is more fully discussed below in *Weaknesses*.

5. The method should identify the true standard (Rating: 3). It is legitimate to say that the Bookmark method identifies the true standard because small standard errors are typically associated with the cut-scores. In the Bookmark method, the use of small groups facilitates the involvement of all panelists in the discussions of items and ratings (Lewis et al., 1998), and feedback is provided throughout the three rounds of the bookmark standard setting. As a result, low standard errors of cut-scores are typically associated with the performance standards. In the implementations listed by Lewis et al. (1998) where cluster sample standard errors were calculated from Round 2 small-group medians, the mean standard errors of cut-scores (in scale standard deviations units) range typically from 0.07 to 0.08 (Lewis et al.). Generally, the patterns of variability among participant judgments can be graphed using median scores from the small groups from Round 1 to Round 3. As shown in Figure 3 (Mitzel et al., 2001), the highest variability happens in the first round when panelists make their first independent ratings, decreases significantly in Round 2, and remains about the same in Round 3. The stability of small-group median scores from Round 2 to Round 3 indicates that the panelists have established a stable perspective as to where to place their bookmarks by Round 3. Although small standard errors indicate that the true cut-score is close by, we can never be absolutely sure where the true cut-score lies.

6. The method should yield decision validity evidence (Rating: 1). The proponents of the Bookmark method have not provided much decision validity evidence. There appears to be little evidence of the accuracy of the decisions based on the cut-scores, that is, the estimates of the probabilities of correct and incorrect classification decisions. However, this is also a weakness associated with every procedure used to set cut-scores. To obtain evidence of decision validity, longitudinal studies are required to investigate how students who scored higher than the cut-score perform and how students who scored lower perform on subsequent tasks. However, the conduct of these studies, particularly when the cut-scores are set for a school-leaving or exit test, is difficult and costly due to the need to follow up on students. Further, the sample of students obtained would probably be restricted due to the inability to locate all students after school graduation. Such follow-up studies are, nevertheless, more feasible when the tests are at the lower grade levels. Performance in the next year or years of school can be used to determine rates of both types of incorrect decisions. Essentially, decision or consequential validity remains a problematic area for the Bookmark procedure and for all other standard-setting procedures. This is not to deny, however, the validity of the Bookmark procedure in that evidential validity is clearly present from the process of setting cut-scores. Serious consideration is given to the background and representativeness of the panelists; panelists are well trained in terms of both their understanding of the assessment and the standard-setting procedure; discussions and feedback are always encouraged, and multiple groups are formed to check the generalizability of standards; both performance data and consequential data are used effectively; and the entire process is clearly documented and performance

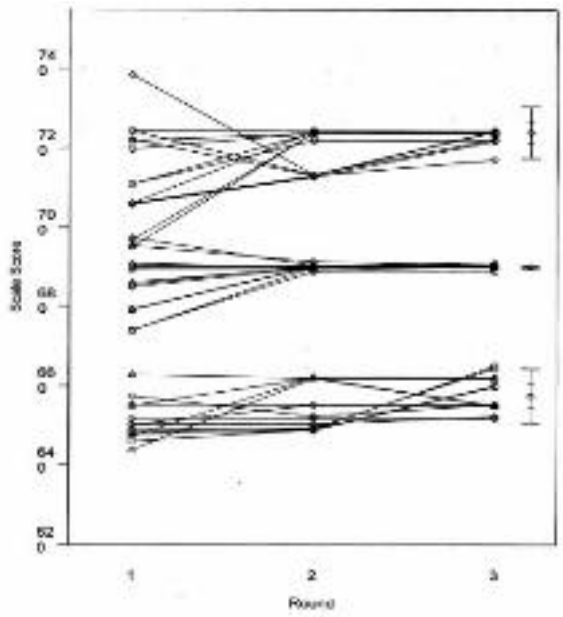

Figure 3. Graphical presentation of participant judgments across rounds.
Note. Taken from Mitzel et al. (2001), p. 257.

standards effectively communicated. One more piece of evidential validity would be the fact that meaningful performance standards are derived once the cut-scores have been set: items before a bookmark represent content that students need to master in order to enter the corresponding performance level. Above all, although evidential validity provides some evidence for validity of the Bookmark procedure, consequential validity is still needed to determine more fully whether the procedure is valid for a particular standard-setting situation.

Practicability

7. The method should be easy to implement (Rating: 3). The Bookmark method is easy to implement. It involves mainly the preparation of the ordered-item booklet, three rounds of bookmark placements, and writing of the performance level descriptors. The manual (Lewis et al., 1999) provides detailed information about the whole process.

8. The method should be easy to compute (Rating: 3). The Bookmark method is easy to compute in terms of the statistical methods used to arrive at the final cut-score. Item ranking using IRT programs is typically completed by psychometricians before the standard-setting panels are formed. After this the calculations of medians, cut-scores, and their associated standard errors can be easily handled using EXCEL.

9. The method should be easy to interpret to laypeople (Rating: 3). The bookmark method is easy to interpret to laypeople. Although IRT might not be easy for laypeople to understand, ranking the items in terms of difficulty and placing bookmarks to divide the items are conceptually acceptable and intuitively sound. The mechanism of IRT is usually not explained to the panelists, but it is made clear that the items are ordered according to their difficulty levels. In a

study of the cognitive experience of the panelists, Dawber and Lewis (2002) concluded that the panelists were able to interpret the definition of mastery correctly and that they understood the Bookmark procedure. Some panelists' understanding improved from Round 1 to Round 2, but most had the same understanding in Round 3 as in Round 2.

10. The method should be credible to laypeople (Rating: 3). The bookmark method is credible to laypeople. Exit surveys with panelists are routinely conducted after standard-setting, and results have been generally positive. In the standard-setting for the Alaska High School Graduation Qualifying Exam (CTB/McGraw-Hill, 2001), for example, over 85% of panelists rated the training and organization of the standard setting good or excellent. Eighty-one percent of panelists were satisfied or very satisfied with their group's bookmark placements, and 88% said that they would defend the cut-scores that were set.

On the whole, the Bookmark procedure fully met nine of Berk's (1986) 10 criteria. In terms of technical adequacy, five of the six criteria are fully met, and all four practicability criteria are fully met. Nevertheless, the technical standard that was not met—providing decision validity evidence—is not only a Bookmark issue. It is typically unsettled in all other standard-setting procedures. Taken together, the Bookmark method is a sound and promising procedure in terms of both technical adequacy and practicability.

Strengths of the Bookmark Procedure

As mentioned above, the Bookmark procedure for setting cut-scores and finalizing performance standards has been widely implemented in the US since its development in 1996. Generally, the success of the Bookmark procedure can be attributed to a number of strengths it possesses: (a) accommodating constructed-response as well as selected-response test items; (b) reducing cognitive complexity for panelists; (c) connecting performance descriptors with the content assessed; (d) promoting better understanding of expected student performance; (e) efficiently accommodating multiple cut-scores; (f) accommodating multiple test forms; (g) reducing the length of time needed to set cut-scores; and (h) reducing standard errors of cut-scores.

Accommodating Constructed-Responses as Well as Selected-Response Test Items

Inclusion of constructed-response item types is necessary in many large-scale tests, especially when writing and complex problem-solving need to be assessed. Traditional standard-setting procedures such as the Angoff and modified Angoff procedures, tend to work better with selected-response items than with constructed-response items (Mitzel et al., 2001). With the Bookmark procedure, constructed-response items appear multiple times in the ordered-item booklet, once for each score point. Thus the constructed-response and selected-response items can be considered together by the panelists.

Reducing Cognitive Complexity for Panelists

The reduction of cognitive complexity required of the panelists is another significant advantage of the Bookmark procedure (Lewis et al., 1998; Mitzel et al., 2001). In item-centered standard-setting procedures such as the Angoff (1971) procedure and its modifications, panelists are first asked to estimate the probability that a hypothetical student at the boundary of a given achievement

level will get a particular item correct, which is deemed an almost impossible cognitive task (Shepard et al., 1993). Then the judgments on individual items are accumulated statistically to form a cut-score. In the Bookmark procedure, items are structured such that the test content can be systematically analyzed so the judgment task is reduced to one of dividing test content between what should be mastered and what need not be mastered for a given performance level. Thus the number of judgments each panelist has to produce is greatly reduced, and so is the cognitive complexity. In addition, by providing panelists with known difficulty information, the Bookmark procedure allows panelists to focus on item content rather than item difficulty (Zieky, 2001), which in turn simplifies the judgmental task required of them.

Connecting Performance Descriptors with the Content of Assessments

As explained above, performance-level descriptors emerge as a final outcome of the Bookmark procedure. After the final cut-score is established, the panelists examine the items before the bookmark and synthesize the content measured by those items. The performance-level descriptors represent a summary of the knowledge, skills, and abilities that students must be able to demonstrate to enter each performance level. According to Mitzel et al. (2001), if performance descriptors are to be used to provide valid guidance to stakeholders of what a student must know and be able to do, the standard-setting procedure should provide a valid way to relate test performance to content mastery. Because the performance descriptors are based on the actual cut-score and student performance, the Bookmark procedure provides defensible performance level descriptors that are tied closely to the content of assessments and what students need to know for each standard (Lewis et al., 1996). It should be noted, however, that writing performance descriptors on the basis of a test requires that valid inferences can be made about student performance. If the test items are not relevant to and representative of the curriculum, the performance descriptors may be biased and flawed from the beginning.

Promoting Better Understanding of Expected Student Performance

The writing of performance descriptors under the Bookmark procedure typically involves examination and synthesis of the content before a bookmark. Consequently, the panelists are more likely to leave the bookmark standard setting with a strong understanding of expected student performance for each performance level. For example, Lewis et al. (1998) conducted a Bookmark standard-setting study that involved 20 panels setting cut-scores in reading, language arts, and mathematics for grades 3 to 10. The findings of this study suggested that panelists using the Bookmark procedure had a systematic understanding of the item pool as a whole, and thus a solid understanding of what the final cut-scores represented in terms of what students in each performance level should know and be able to do. In addition, Bookmark panelists frequently commented on how instruction would improve if every teacher could go through the same process (Lewis et al). Apparently, writing performance descriptors after standard-setting allows panelists to understand better how the assessment is related to the content standards, curriculum, and instruction to which the assessment is referenced.

Efficiently Accommodating Multiple Cut-Scores

When there is more than one cut-score, panelists using a modified Angoff procedure need to judge the probability that a hypothetical student at each of the boundaries of the series of achievement levels will get a particular item correct. That is, for each cut-score, panelists need to make new judgments on every item. In contrast, panelists using the Bookmark procedure can set multiple cut-scores efficiently one after another using the ordered-item booklet. Setting the cut-score of the next level, for example, means simply reviewing the items after the first bookmark, which is efficient in terms of both labor and time.

Accommodating Multiple Test Forms

As an IRT-based approach, the Bookmark procedure enjoys advantages that IRT brings. One advantage is the ability to accommodate multiple test forms in one standard setting. If multiple tests sampled from a common domain can be placed on a common scale using IRT methods, all the items can then be ordered in one booklet. The ordered-item booklet can span up to 110 score points (Mitzel et al., 2001), which makes it possible to combine more than one test. Therefore, the ability to present a content domain that is more representative than a single test form is viewed as another strength of the Bookmark procedure (Mitzel et al.).

Reducing the Length of Time Needed to Set Cut-Scores

Buckendahl, Smith, Impara, and Plake (2002) compared a modified Angoff procedure and a modified Bookmark procedure when setting cut-scores for a grade 7 mathematics assessment (selected-response items only). Two panels consisting of 12 teachers for the Angoff procedure and 11 teachers for the bookmark were established to set the cut-scores. The Angoff group was asked to conceptualize a specific barely proficient student they had taught and then indicate for each item whether the student they had in mind would answer the item correctly or not. After seeing the performance data, the judges were asked to make a second estimate of each item, whether the same or different from their first estimate. The recommended cut-score based on the second estimates was calculated by summing the number of correct items for each teacher and then averaging the values across the teachers. For the Bookmark procedure, the items were first ordered from the easiest to the most difficult, using p values estimated from a pilot test (rather than b parameters in IRT models). The judges were then asked to conceptualize a specific barely proficient student they had taught, start with the easiest item, and move through the booklet until they found the place where their barely proficient student would probably get all items up to that point correct and all items after that point incorrect. At that point in the booklet, the judges placed their bookmarks. After the presentation of the performance data, each judge was asked to make a second bookmark placement. The final cut-score based on the second round results was calculated by summing the number of items up to the bookmark for each teacher and then averaging the values across the teachers. This study reported similar levels of confidence in the passing score and comfort in the process followed between the two groups. In agreement with Lewis et al. (1996), Buckendahl et al. suggested that the Bookmark procedure might be more efficient in terms of

the length of time it took for the panelists to make their bookmark placements, which was shorter than the time required to complete the yes/no Angoff procedure.

Reducing Standard Errors of Cut-Scores

Buckendahl et al. (2002) also compared the mean cut-scores obtained from the two methods for the 69-item test, and the difference was found to be small (33.42 for the Angoff and 35.64 for the Bookmark). Nevertheless, the Bookmark method produced a lower standard deviation of the cut-scores (10.96 for the Angoff and 8.66 for the Bookmark), which indicated better interjudge agreement. Therefore, despite of the use of modified Angoff and modified Bookmark methods, Buckendahl et al. provide some evidence of the Bookmark procedure's advantage in both efficiency and accuracy.

Weaknesses of the Bookmark Procedure

Despite its strengths, the Bookmark procedure has some potential weaknesses. These include the choice of response probability, item disordinality, exclusion of important factors other than the difficulty of the items, and restrictions of the IRT models.

Choice of Response Probability

In the Bookmark procedure, items are ordered according to their locations on the ability scale when the RP is set to 0.67. In spite of the support from the research of Huynh (1998), the use of 0.67 as the RP is often questioned. Although the choice of RP tends to have a small effect on the ordering of items in terms of difficulty (Egan, 2001), the cut-scores may be manipulated by changing the RP value (Kolstad, 1996). Mitzel et al. (2001) concur that one of the unresolved issues with the Bookmark procedure is the ordering of the items, because items can be ordered slightly differently using different RP values. In the study of Beretvas (2004), the use of different RP values (1/2, 2/3, 4/5) resulted in somewhat different ranking of bookmark difficulty locations, with correlations between difficulty orderings for pairs of RP values ranging from 0.93 to 0.99. Taking the items in Figure 2, for example, if a lower RP (e.g., 0.50) were set, the order of items 3 and 4 in the ordered-item booklet would be switched. Because bookmark placement depends on the ordering of items, different RP values may thus produce somewhat different cut-scores, especially when the ordering of items near the cut-points is affected.

Item Disordinality

Item disordinality refers to "the disagreement among judges on the ordering of the items in the booklet" (Skaggs & Tessema, 2001, p. 2). According to Lewis and Green (1997), item disordinality is an outstanding issue in virtually all applications of the Bookmark method. Typically, panelists do not agree on how items are ordered in the booklet because they may have different local curricula and/or they are not able to estimate item difficulty accurately (Lewis & Green). In the standard setting for the Alaska High School Graduation Qualifying Exam (CTB/McGraw-Hill, 2001), for example, 15% of panelists generally disagreed or somewhat disagreed with the item ordering, and another 14% remained neutral for this question. As a result of item disordinality, the variability of the cut-scores among the panelists may increase. This is especially

a problem when item disordinality occurs near the cut-points. To resolve disordinality disagreement, Lewis and Green recommended a thorough discussion among the panelists of what each item measures and what makes it more difficult than the preceding item. Nevertheless, these discussions did not completely resolve the disordinality disagreement in Skaggs and Tessema's study.

Exclusion of Important Factors Other than the Difficulty Parameter

In the Bookmark procedure, the items used for standard setting are ordered simply according to their difficulties. Although this reduces the cognitive load on the part of the panelists, it "does not allow participants to distinguish purposefully among the items above the bookmark, or among the items below the bookmark on the basis of importance, curricular relevance, or necessity for performance on the job" (Zieky, 2001, p. 35). Depending on the types of assessments, however, these factors may be important considerations in setting cut-scores. Taking mathematics tests for example, items measuring problem-solving skills may be more important than items measuring knowledge only. When ranked according to difficulty only, these problem-solving items, which usually have higher difficulty, will be placed more toward the back of the booklet and thus will be more likely to be excluded from the content requirement for a given performance level. In certain assessment settings, difficulty should not be the only factor used to rank the items; importance or necessity for performance on the job should also be taken into account.

Restrictions of the IRT Models

As an IRT-based method, the Bookmark procedure is restricted in some ways by the assumptions of the IRT models. That is, the use of the Bookmark procedure is somewhat conditional on the satisfaction of assumptions underlying the development and use of IRT. These assumptions include essential unidimensionality (Stout, 1987), local independence, and non-speededness. If any of these assumptions is not satisfactorily met, the robustness of setting cut-scores using unidimensional models should be questioned.

Discussions and Conclusions

The Bookmark procedure was developed to address the perceived problems with the Angoff procedure and its modified variations—the most popular procedures for setting cut-scores. When evaluated using Berk's (1986) consumer's guide to setting performance standards, five of the six technical criteria are fully met. The Bookmark method yields appropriate classification information; identifies the true cut-score; is sensitive to examinee performance, instruction, or training; and is statistically sound. The problematic area is the lack of decision validity evidence. In terms of practicability, all four criteria are fully met. The Bookmark procedure is relatively easy to implement, compute, and interpret to laypeople.

Generally, the strengths of the Bookmark procedure lie mainly in its accommodation of both constructed-response and selected-response test items, its reduction of cognitive complexity, its connection of performance descriptors with the content of assessments, its promotion of better understanding of expected student performance, and its accommodation of multiple cut-scores and multiple test forms. When compared with a modified Angoff procedure,

the Bookmark method may be more efficient in terms of the length of time it takes for judges to make their bookmark placements, and the standard deviation of its mean cut-score is also lower (Buckendahl et al., 2002).

When it comes to weaknesses of the Bookmark procedure, the choice and understanding of the response probability remain outstanding issues. Items can be ordered differently in an ordered-item booklet using values other than the typical 0.67. Further, item disordinality may affect the generalizability of the cut-scores. Although there is evidence (Mercado, Egan, & Brandstrom, 2003; Dawber & Lewis, 2002) that suggests that bookmark participants are able to understand the application of the RP criterion, it is not clear how well panelists perceive, internalize, and use the response probability in the process of cut-score-setting, and how this in turn affects their cut-score placements (Mitzel et al., 2001). In addition, the restrictions of IRT assumptions may be a problem in practical applications of the Bookmark procedure. Finally, the research of consequential validity remains a weak area for the Bookmark procedure. However, as pointed out above, evidence of consequential validity for other standard-setting procedures is also rare. The common difficulties in validating standard-setting procedures apply in the Bookmark procedure. When methods of convergent validity are used to validate standard-setting results, one consistent finding is that varying standard-setting procedures produce varying results (Mitzel et al., 2001). Difficulty in collecting data for external validity checks inevitably contributes to the rarity of such evidence. In fact despite the growth of the number of standard-setting methods and their variations, it remains "difficult to set standards and even more difficult to validate standards" (Kane, 2001, p. 54).

An additional concern is that the use of group discussions, normative information, and impact data in standard-setting procedures such as the Bookmark "has the primary effect of regressing what might result from any particular standard-setting procedure toward what is" (Cizek, 2001, p. 11). As a result, among the four major uses of performance standards, exhortation, exemplification, accountability, certification, and recertification (Linn, 1994), the goal of exhortation may not be reached. That is, instead of motivating teachers and students to greater levels of achievement, the Bookmark procedure tends to reflect the current achievement and therefore its use for accountability purposes.

To summarize, the strengths of the Bookmark procedure clearly outweigh its weaknesses. The Bookmark procedure remains a promising procedure with its own characteristics and advantages. More research will certainly benefit this relatively new method in standard-setting, especially studies in consequential validity, cognitive processing, and the criterion of response probability.

Acknowledgments

An earlier version of this article was presented at the Annual Meeting of the Canadian Society for the Study of Education, Halifax, Canada, 2003. I thank W. Todd Rogers for his constructive and insightful feedback during the development of this article. Thanks also go to the two anonymous reviewers for their helpful comments on the earlier version.

References

Angoff, W.H. (1971). Scale, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303-310.

Beretvas, S.N. (2004). Comparison of Bookmark difficulty locations under different item response models. *Applied Psychological Measurement, 28*(1), 25-47.

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

Buckendahl, C.W., Smith, R.W., Impara, J.C., & Plake, B.S. (2002). A comparison of the Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*, 253-263.

Burket, G.R. (1991). *PARDUX [computer program]*. Unpublished.

Cizek, G. (2001). 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.

Cochran, W.G. (1963). *Sampling techniques* (2nd ed.). New York: John Wiley & Sons.

CTB/McGraw-Hill (2001). *Alaska high school graduation qualifying exam: HSGQE standard setting technical report*. Retrieved January 3, 2005, from: http://www.eed.state.ak.us/tls/assessment/HSGQE/HSGQEstanset_techrprt1.pdf

Dawber, T., & Lewis, D.M. (2002). *The cognitive experience of bookmark standard setting participants*. Paper presented at the Annual meeting of the American Educational Research Association, New Orleans.

Egan, K.L. (2001). *Validity and defensibility of cutscores established by the Bookmark standard setting method*. Paper presented at the 2001 Council of Chief State School Officers Conference on Large-Scale Assessment, Houston.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23*(19), 35-56.

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

Kolstad, A.J. (1996). *1992 national adult litercy survey: Literacy levels and the 80 percent response probability convention*. Washington, DC: National Centre for Education Statistics.

Lewis, D.M., & Green, D.R. (1997). *The validity of performance level descriptors*. Paper presented at the Council of Chief State School Officers National Conference on Large-scale Assessment, Phoenix.

Lewis, D.M., Green, D.R., Mitzel, H.C., Baum, K., & Patz, R.J. (1998). *The bookmark standard setting procedure: Methodology and Recent Implementations*. Paper presented at the National Council for Measurement in Education annual meeting, San Diego.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996). 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.

Lewis, D.M., Mitzel, H.C., Green, D.R., & Patz, R.J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

Linn, R.L. (1994). *The likely impact of performance standards as a function of uses: From rhetoric to sanctions*. Paper presented at the National Centre for Education Statistics and National Assessment Governing Board joint conference on Standard-Setting for Large-Scale Assessments, Washington, DC.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. New York: Elbaum.

Lord, F.M., Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mercado, R.L., Egan, K.L., & Brandstrom, A.J. (2003). *The response probability criterion in the bookmark standard setting procedure*. Paper presented at the National Council for Measurement in Education annual meeting, Chicago.

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.

Muraki, E., & Bock, R.D. (1991). *PARSCALE: Parameter scaling of rating data* [computer program]. Chicago, IL: Scientific Software.
Ricker, K.L. (in press). Setting cut-scores: Critical review of Angoff and modified-Angoff methods. *Alberta Journal of Educational Research.*
Shepard, L., Glaser, R., & Bohrnstedt, G. (Eds.). (1993). *Setting performance standards for student achievement*. Stanford, CA: National Academy of Education.
Skaggs, G., & Tessema, A. (2001). *Item disordinality with the bookmark standard setting procedure*. Paper presented at the National Council for Measurement in Education annual meeting, Seattle.
Stout, W.F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 289-617.
Thissen, D. (1999). *MULTILOG* [computer program]. Chicago, IL: Scientific Software.
Wisconsin Department of Public Instruction. (2003). Bookmark standard setting overview. Retrieved January 3, 2005, from: http://www.dpi.state.wi.us/oea/ctbbkmrk03.html
Zieky, M.J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.