

The Bookmark Standard Setting Procedure:

Strengths and Weaknesses

Jie Lin

The Centre for Research in Applied Measurement and Evaluation

The University of Alberta

Edmonton, Alberta, Canada

Abstract

The Bookmark standard setting procedure was developed to address the perceived problems with the most popular method for setting cutscores, the Angoff procedure (Angoff, 1971). The purpose of the present paper is to review the Bookmark procedure and then evaluate it in terms of Berk's (1986) criteria for evaluating standard-setting methods. Finally, the strengths and weaknesses of the Bookmark procedure are critically examined and discussed. In conclusion, the strengths of the Bookmark method are that it (a) accommodates constructed-response as well as selected-response test items; (b) efficiently accommodates multiple cutscores and multiple test forms; and (c) reduces cognitive complexity for panelists. Despite of unsolved issues like the choice and understanding of the response probability, the Bookmark method is a promising standard setting procedure with its own characteristics and advantages.

The Bookmark Standard Setting Procedure:
Strengths and Weaknesses

Standard setting 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 cutscores corresponding to a specified performance level, the Angoff procedure (Angoff, 1971) is often considered “the industry standard” (Zieky, 2001). However, dissatisfaction with this procedure has been growing. 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 evaluation of all the items for every standard level and the production of numerous p values can be tedious (Mitzel, Lewis, Patz, & Green, 2001). Poor accuracy has also been found in the judgments of item difficulty magnitude. Typically, panelists tend to overestimate performance on difficult items and underestimate performance on easy items (Bejar, 1983). Last, it still remains an open question whether teachers 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 be, instead, “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 critical review of Angoff and modified Angoff procedures, see Ricker (2004).

To address the perceived problems of the Angoff procedure, Lewis, Mitzel, and Green (1996) developed the bookmark standard setting procedure. The Bookmark procedure aims to (a) simultaneously accommodate selected-response and constructed-response test formats, (b) simplify the cognitive complexity required of standard setting judges, (c) connect the judgment task of setting cutscores 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 standard setting procedure to set cutscores on their 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 (e.g. Lewis, Green, Mitzel, Baum, & Patz, 1998;

Lewis, Mitzel, Green, & Patz, 1999; Mitzel et al., 2001). The intent of the present paper is, therefore, to provide an independent review of the Bookmark procedure, evaluate it in terms of Berk's (1986) criteria for evaluating standard-setting methods, and critically examine its strengths and weaknesses.

Basic Assumptions of the Bookmark Standard Setting 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), which displays the relationship between the proficiency of an examinee and the probability of 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.

 Insert Figure 1 about here.

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 cutscore correctly. The use of 0.67 as the response probability (RP) has been supported by the research of Huynh (1998). Huynh (1998) showed that for the 3PL model, the item information function is maximized at θ for which $P(\theta) = (c + 2) / 3$. When guessing is factored out ($c = 0$), the value of RP equals $2/3$.

The three-parameter logistic (3PL) model (Lord & Novick, 1968; Lord, 1980) 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, 1991), 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 cutscore for every performance level.

Overview of the Bookmark Standard Setting Procedure

The Bookmark Standard Setting 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 key 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 spot on the ability scale for which students have a 0.67 probability of achieving that score point or higher. 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 conceptualization of what the test measures, as well as to serve as a vehicle to make cutscore judgments” (p. 7).

 Insert Figure 2 about here.

Along with the ordered item booklets, item mapping rating forms are also provided as a guide to the booklets. The rating forms list all the items in the same 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 the cutscores they are considering for each round (Lewis et al., 1998).

Panel Composition and Training

Each rating panel should consist of at least 18 panelists for each grade/content area (Lewis et al., 1998), although 24 is 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 group consists of 5 to 7 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 the standard setting task is also emphasized to the 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. For a typical Bookmark conference agenda, see Table 1.

 Insert Table 1 about here.

Setting Cutscores

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 then asked to discuss and determine the content that students should master for placement into a given performance level.

Their independent judgments of cutscores 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 of less than 0.67.

Round 2.

The first activity in Round 2 involves having each member place bookmarks in his/her ordered item booklet where each of the other panelists in their small group made their bookmark placement. For a group of 6 people, each panelist's ordered booklet will have 6 bookmarks for each cut point. Discussions are then focused on the items between the first and last bookmarks for each performance level. Upon completion of this discussion, the panelists then 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 cutscore from Round 2. With this information of how students actually performed, the panelists discuss the bookmarks in the large group and then make their Round 3 independent judgments of where to place the bookmarks. The median for the large group is considered to be the final cut-point for a given performance level.

Definition of Cutscores

As mentioned before, 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 make them easier for the public to understand, the theta scores are linearly 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 cutscore for that particular level.

Finalizing Performance Standards

Based on the final cutscores set, performance level descriptors are then written by the panelists. Performance descriptors describe the specific knowledge, skills, and abilities held by students at a given performance level. Items prior to 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 then synthesized to formulate the descriptors for this performance level. Performance level descriptors thus become a natural extension of the cutscore setting procedure.

Evaluation of the Bookmark Procedure Using Berk's (1986) 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" (Berk, 1986, p. 140). However, a technically defensible procedure does not necessarily warrant a feasible procedure in that the procedure may be very 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 cutscores produced by the Bookmark method permit dichotomous classification decisions at each cutscore point. For example, students with scale scores higher than the cutscore for the proficient level are considered "proficient", while those who score lower than the cutscore are classified "non-proficient".

2) *The method should be sensitive to examinee performance. (Rating: 3)*

The Bookmark method is sensitive to examinee performance. 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 the setting of the final cutscores.

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 perform poorly on these items. That is, the difficulty parameters of these items would be relatively high at the response probability of 0.67. Therefore, in the ordered item booklet, these items would be positioned more towards the back rather than the front. The bookmarks, then, would be more likely to be set prior to them, meaning that these uncovered items would not be included in the performance level descriptors and therefore not be considered something the students needed to master to achieve a given standard. Likewise, items addressing the content covered well in the classroom would be answered successfully by the majority, achieve lower difficulty values, and therefore would be more likely classified as content a student has to master to achieve a certain level. In this way, the instruction or training received by examinees is reflected in the order of items, which then influences the setting of cutscores.

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 cutscores 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 Bookmark procedure. As for the calculation of the standard error, the cluster sample standard error (Cochran, 1963, p. 210) was calculated from the Round 2 small group medians. Since 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 cutscores 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 cutscores and standard errors. The only concern, however, would be the selection of response probability, that is, to what extent the cutscore may be manipulated by changing the response probability value. This concern is more fully discussed in the section of weaknesses of the Bookmark procedure.

5) *The method should identify the true standard. (Rating: 3)*

The Bookmark method identifies the true cutscore. Based on the IRT, the bookmark approach identifies the cutscore on the theta (θ) score scale rather than an observed score scale. 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 cutscores 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 standard errors of cutscores (in scale standard deviations units) range typically from 0.07 to 0.08. 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 (Lewis et al., 1998), the highest variability happens in the first round, when panelists make their first independent ratings, and decreases significantly in Round 2, and remains about the same in Round 3. The stability of individual cutscores from Round 2 to Round 3 indicates that the panelists have developed a stable perspective as to where to place their bookmarks by Round 3. While small standard errors indicate that the “true” cutscore is close by, we can never be sure where the true cutscore lies.

 Insert Figure 3 about here.

6) *The method should yield decision validity evidence. (Rating: 2)*

The proponents of the Bookmark method have not provided much decision validity evidence. There appears to be no evidence of the accuracy of the decisions based on the cutscores, that is, the estimates of the probabilities of correct and incorrect classification decisions. However, this is also a weakness associated with most cutscore setting procedures. To obtain evidence of decision validity, longitudinal studies are required to investigate how students who scored higher than the cutscore perform and how students who scored lower than perform on subsequent tasks. However, the conduct of these studies, particularly when the cutscores are set for a school leaving or exit test, is difficult and costly due to the need to follow-up students. Further, the sample of students obtained would likely 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. This is not to deny, however, the validity of the Bookmark procedure in that evidential validity is clearly present from the process of setting

cutscores. 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 standards are effectively communicated. Although evidential validity alone provides some evidence for validity of the Bookmark procedure, consequential validity is still needed to more fully determine 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 process.

8) *The method should be easy to compute. (Rating: 3)*

The Bookmark method is relatively easy to compute. Psychometricians are needed to run IRT computer programs in order to rank the items in the ordered item booklet. After that, the calculations of medians, cutscores and their associated standard errors can be easily handled using EXCEL.

9) *The method should be easy to interpret to laypeople, and*

10) *The method should be credible to laypeople. (Rating: 3)*

The bookmark method is relatively easy to interpret to laypeople, and relatively credible 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. In actual standard setting, 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.

On the whole, the Bookmark procedure fully met nine of Berk's ten criteria. The standard that was partially met --- providing decision validity evidence --- is also unsettled in the case of

other cutscore setting procedures. Specifically, in terms of technical adequacy five of the six criteria are fully met, and all four practicability criteria are fully met. Taken together, the Bookmark method is a relatively sound procedure in terms of both technical adequacy and practicability.

Strengths of the Bookmark Procedure

As mentioned earlier, the Bookmark standard setting procedure has been widely implemented in the United States 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 of assessments; (d) promoting better understanding of expected student performance; (e) efficiently accommodating multiple cutscores; (f) accommodating multiple test forms; and (g) time efficiency and low standard error of the cutscores.

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 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-centred standard setting procedures such as the modified Angoff, 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, Glaser, & Bohrnstedt, 1993). Then the judgments on individual items are accumulated statistically to form a cutscore. In the Bookmark procedure, items are structured in a way that the test content can be systematically analyzed so the judgment task is reduced to one of dividing test content between

that which should be mastered and that which 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 the panelists.

Connecting Performance Descriptors with the Content of Assessments

As explained earlier, performance level descriptors emerge as a final outcome of the Bookmark procedure. After the final cutscore 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. Since the performance descriptors are based on the actual cutscore 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, Mitzel, & Green, 1996). It should be noted, nevertheless, 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 or representative of the curriculum, the performance descriptors may be biased and flawed from the very 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 cutscores 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 more systematic understanding of the item pool as a whole, and thus a better understanding of what the final cutscores 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., 1998). Apparently, writing performance descriptors after standard setting allows panelists to better understand how assessment is related to content standards, curriculum, and instruction.

Efficiently Accommodating Multiple Cutscores

When there is more than one cutscore, 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 cutscore, panelists need to make new judgments on every item. In contrast, panelists using the Bookmark procedure can set multiple cutscores efficiently one after another, using the ordered item booklet. Setting the cutscore of the next higher level, for example, means simply reviewing the items after the first bookmark, which is very efficient in terms of both labour and time.

Accommodating Multiple Test Forms

As an IRT-based approach, the Bookmark procedure enjoys advantages that IRT brings about. 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 80 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.).

Time Efficiency and Lower Standard Error

Buckendahl et al. (2000) compared a modified Angoff procedure and a modified Bookmark procedure when setting cutscores 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 cutscores. The Angoff group were 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 same or different from their first estimate. The recommended cutscore, based on the second estimates, was calculated by summing the number of “right” 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 find 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 cutscore, 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. also 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. When the mean cutscores obtained from the two methods were compared for the 69-item test, the difference was small (33.42 for the Angoff and 35.64 for the bookmark). However, the Bookmark method produced a lower standard deviation of the cutscores (10.96 for the Angoff and 8.66 for the Bookmark), which indicated better inter-judge agreement. Therefore, despite of the use of modified Angoff and modified Bookmark methods, Buckendahl et al. provide some evidence on 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 parameter, 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 response probability (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 cutscores may be potentially manipulated by changing the RP value (Kolstad, 1996). Mitzel et al. (2001) agreed 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 (other than the

typical 0.67). 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. Since bookmark placement depends on the ordering of items, different RP values may thus produce somewhat different cutscores, 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 the way 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, cited in Skaggs & Tessema, 2001). As a result, the variability of the cutscores among the panelists may increase, and standard error of the final cutscore will increase accordingly. This is especially a problem when item disordinality occurs near the cut points. To resolve disordinality disagreement, Lewis and Green (1997) 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 (2001).

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. While 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 cutscores. 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 towards the back of the booklet, and thus more likely be excluded from the content requirement for a given performance level. In other words, 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 cutscores 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 cutscores. When evaluated using Berk's (1986) consumer's guide to setting performance standards, five out of the six technical criteria are fully met. The Bookmark method yields appropriate classification information, identifies the true cutscore, is sensitive to examinee performance, instruction or training, and is statistically sound. The problematic area is the 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 mainly lie 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 cutscores and multiple test forms. When compared with the Angoff procedure (1971), the Bookmark method may be more efficient in terms of the length of time for judges to make their bookmark placements, and the standard deviation of its mean cutscore is also lower (Buckendahl et al., 2000).

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 cutscores. Although there is evidence (Mercado, Egan, & Brandstrom, 2003; Dawber & Lewis, 2002) suggesting that bookmark participants are able to understand the application of the RP criterion, it is not very clear how well panelists perceive,

internalize, and use the response probability in the process of cutscore setting, and how that in turn affects their cutscore placements (Mitzel et al., 2001). In addition, the restrictions of IRT assumptions may be a problem in practical applications of the Bookmark procedure. Finally, a common difficulty in validating standard setting procedures also applies in the Bookmark procedure---the multiple methods in convergent validity designs produce non-consistent results (Mitzel et al., 2001). As discussed previously in Berk's evaluation criteria, validity research remains a weak area in the application of the Bookmark procedure.

An additional concern is that the use of group discussions, normative information, and impact data in cutscore 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 sum up, the strengths of the Bookmark method 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 validity, cognitive processing, and the criterion of response probability.

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.
- Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7, 303-310.
- 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. (2000). *A comparison of the Angoff and Bookmark standard setting methods*. Paper presented at the Annual meeting of the Mid-Western Educational Research Association, Chicago, IL.
- Bucket, G. R. (1991). *PARDEX* [computer program]. Unpublished.
- Cizek, G. (2001). Conjectures on the rise and fall 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.
- 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, LA.
- 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.
- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Erlbaum.
- 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.
- Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice*, 10(2), 3-6, 10.
- 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 Literacy Survey: Literacy levels and the 80 percent response probability convention*. National Centre for Education Statistics: Washington, DC.
- 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, AZ.
- 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, CA.
- 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, IL.
- 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. (2004). Setting cutscores: critical review of Angoff and modified-Angoff methods. *Alberta Journal of Educational Measurement*.
- Shepard, L. (1994). *Implications for standard setting of the NAE evaluation of NAEP achievement levels*. Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, Washington, DC.
- 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 disorderliness with the bookmark standard setting procedure*. Paper presented at the National Council for Measurement in Education annual meeting, Seattle, WA.
- 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 May 10, 2003, from <http://www.dpi.state.wi.us/oea/profdesc.html>
- Wisconsin Department of Public Instruction (2003). *Proficiency score standards*. Retrieved May 13, 2003, 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.

Table 1

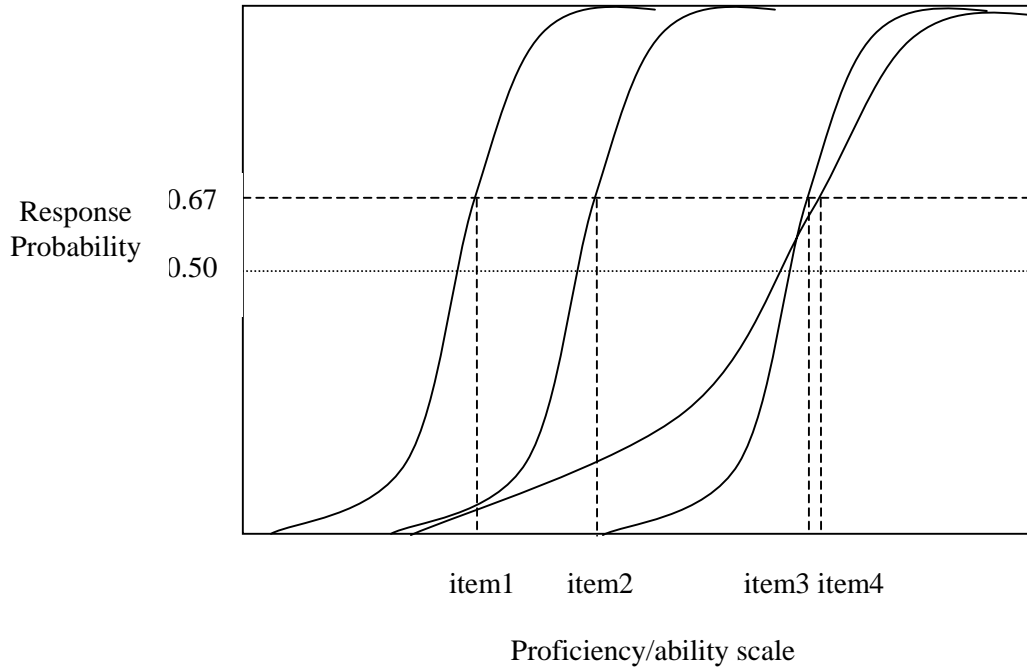
Typical Bookmark Conference Agenda

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

Note: Adapted from Mitzel, Lewis, Patz, & Green, (2001), p. 253.

Figure 1

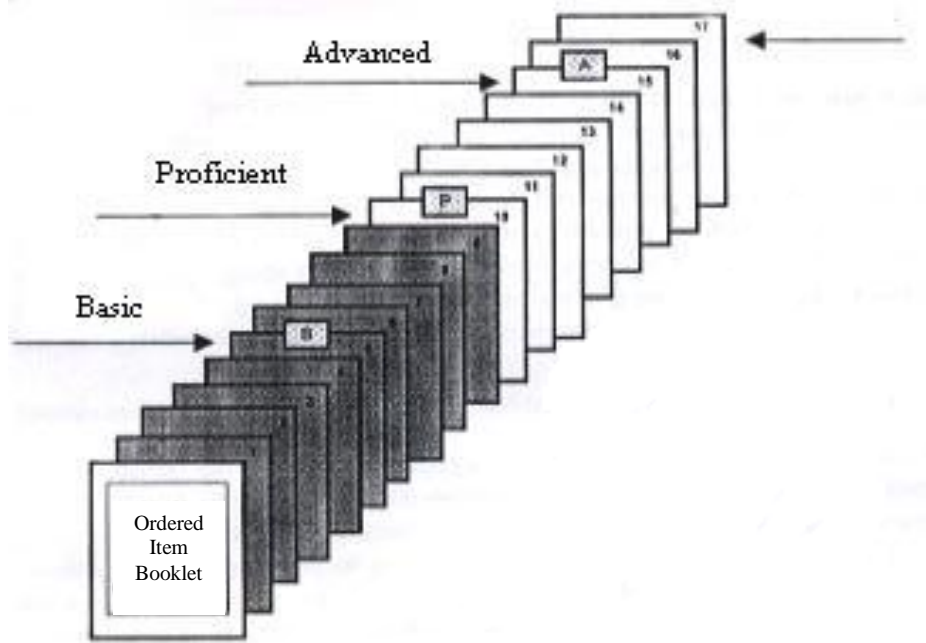
Item Characteristic Curves (ICCs) for SR Items Mapped at RP = 0.67



Note. Adapted From Mitzel, Lewis, Patz, & Green (2001), p. 261.

Figure 2

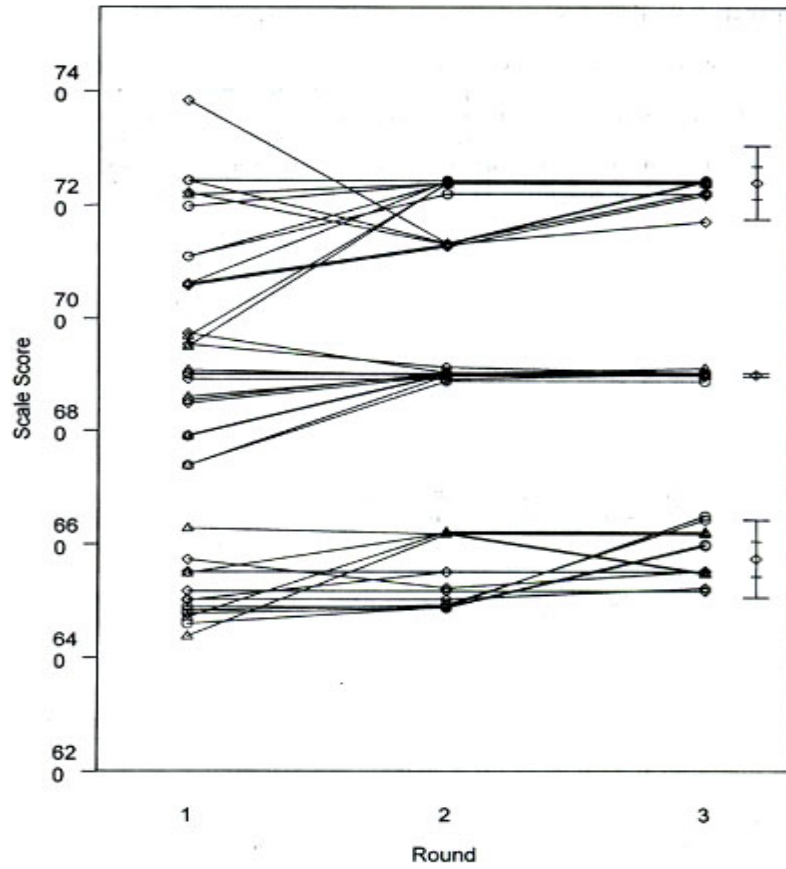
Illustration of Ordered Item Booklet for the Bookmark Procedure



Note. Adapted From Mitzel, Lewis, Patz, & Green (2001), p. 253.

Figure 3

Graphical Presentation of Participant Judgments Across Rounds



Note. Taken from Mitzel et al. (2001), p. 257.