

Running head: COGNITIVE EXPERIENCE OF BOOKMARK PARTICIPANTS

The Cognitive Experience of Bookmark Standard Setting Participants

Teresa Dawber

University of Alberta

Daniel M. Lewis

CTB/McGraw-Hill

Abstract

The purposes of the study were to investigate participants' understanding of the Bookmark procedure, item selection strategies, and factors influencing judgments. Data were collected at two state standard settings. Participants completed surveys and table leaders completed think-aloud procedures following each round. Survey results indicated understanding improved from Round 1 to Round 2, and remained constant in Round 3. Strategies for item selection included identifying an interval of items or a single item with consideration to the preceding items. In Round 1, participants relied on their experiences working with students and understanding of state content standards to select their bookmarks. In Rounds 2 and 3, participants relied on the opinions expressed by their peers. Results of the think-aloud procedure mirrored the survey results.

The Cognitive Experience of Bookmark Standard Setting Participants

The Bookmark standard setting procedure (Lewis, Mitzel, & Green, 1996) has been widely implemented since its inception. Bookmark participants express their judgments by placing bookmarks in an ordered item booklet that reflect their expectations for student achievement at each performance level. A purported advantage of the Bookmark procedure is that it simplifies the cognitive task required of participants (Lewis, Green, Mitzel, Baum, & Patz, 1998).

Although the Bookmark procedure is commonly used, little is known about the cognitive experiences of the participants; that is, their understanding of the Bookmark procedure, how they decide where to place their bookmarks, and what influences participants to change their judgments from round to round. It is important to document these aspects of the Bookmark experience. First, it is fundamental to the procedural evidence supporting the interpretations of the standards set using the Bookmark procedure. Validity is not a property of the standard setting method, but rather the interpretations derived from the test scores, and these interpretations are inferred to be valid if convincing evidence supports them (AERA, APA, NCME, 1985, 1999). The aim of validation is to present supporting evidence that the cutscores represent the intended performance standards and that the performance standards are appropriate, reflecting the goals of the decision process. Procedural evidence is crucial to evaluate the appropriateness of the performance standards because few empirical checks on the performance standards are available. We may have confidence in the standards if they have been derived by consensus, and have been set by persons who are unbiased, knowledgeable about the purpose of the standard setting, and understand the process they

are using (Kane, 2001). Although procedural evidence cannot verify the validity of cutscores and their performance standards (Messick, 1989), procedural evidence is often considered adequate to provide basic support for the performance standards and cutscores in the absence of conflicting evidence suggesting the performance standards or cutscores are inappropriate (Kane, 2001).

Second, understanding the process is consistent with the call by measurement specialists to investigate the unique information provided by the various standard setting methods (Lewis, 2001; Mitzel, Lewis, Patz, & Green, 2001). Comparative studies of standard setting methods have found that different methods yield different cutscores (Jaeger, 1989). Mitzel et al. recommend that standard setting research investigate what is being judged, the responses requested by panelists, and the cognitive processes that panelists use to formulate their responses. Although the standard setting tasks outlined for participants may appear straightforward, discussions with participants during standard settings reveal more complex cognitive processes than might be anticipated (Raymond & Reid, 2001). By investigating participants' rationales for making their judgments we may better understand the information provided by a Bookmark standard setting.

Consequently, given the need for procedural evidence for the Bookmark procedure, the following questions were addressed in the present study:

1. What is participants' initial understanding of the Bookmark procedure and does this understanding evolve from round to round?
2. Do participants make Tc 0.390v.essatia e usl thtem,ia eme card ing htems,ard

3. What factors do participants consider when making their initial bookmark placements and do these factors change from round to round?

Method

The Bookmark Procedure

The Bookmark standard setting procedure capitalizes on item mapping procedures (Lewis et al., 1996). Items are ordered by their scale locations, which are based on item response theory (IRT) calibrations (Lewis, Mitzel, & Green, 1996). The response probability of .67 is typically used to order the items according to their scale locations to form an ordered item booklet. The definition of mastery of a performance level is given as: We say that a group of like students has demonstrated mastery of the content represented by an item if at least $2/3$ (67/100) of the students in the group can be expected to respond successfully to the item.

The structure of the Bookmark procedure typically involves three rounds. Participants work in small groups through the first two rounds. In Round 1, participants complete the operational test and study the ordered item booklet (i.e., discuss what each item measures and why the item is more difficult than the previous items in the ordered item booklet), and make individual and independent bookmark placements. In Round 2, participants review and discuss the rationales for the Round 1 judgments within their small groups, and make new judgments. In Round 3, participants discuss their Round 2 judgments as a large single group and make a final set of judgments.

Participants

Data were collected from 69 participants during two state standard settings held in 2001. The purpose of both standard settings was to set cut scores for high school exams.

The first standard setting (SS1) was for a math exam, and the second standard setting (SS2) was for math and science exams (SS2m, SS2s respectively). All three exams were composed of multiple-choice questions only. The three standard setting committees each were composed of 23 participants. Participants were selected by the sponsoring departments of education and were intended to be representative of the state in terms of geographic location, socioeconomic status, ethnicity, gender, and community composition (i.e., urban, rural, suburban).

There were three small groups composed of seven or eight members in SS1, and four small groups composed of five or six members in SS2. Each small group had a table leader, a participant who received additional training and facilitated portions of the process at each table. There was one notable difference between the two standard settings. Impact data (percent of students falling within each performance level category based on group judgments) was presented prior to the final two rounds of judgments at SS2. No impact data was presented at SS1. The decision to include impact data is a policy decision made by the sponsoring department of education, and is not fundamental to the Bookmark procedure.

Demographic information for SS1 was obtained from the state department of education. Information for 21 of the 23 participants was obtained. There were 14 (67%) female and 7 (33%) male participants. Five participants (24%) were of racial or ethnic minority status. Demographic information for the SS2 participants was obtained from an evaluation form completed at the end of the standard setting. Twenty-two participants from each committee completed the demographic evaluation form. On the math committee, there were slightly more teachers ($n = 12$) than non-teachers ($n = 10$),

whereas there were equal numbers ($n = 11$) of teachers and non-teachers on the science committee. The mean number of years participants have been working in their current profession were 17.43 and 22.68 for SS2m and SS2s respectively. Two participants on SS2m and three participants on SS2s identified themselves as being of racial or ethnic minority status. There were equal numbers of males and females on SS2m ($n = 11$), but twice as many females ($n = 16$) than males ($n = 8$) on SS2s.

Data Collection

Although more than one cutscore was set at each standard setting, only the proficient, or passing, cutscore was addressed in this research. The proficient cutscore was chosen because it is the primary category; that is, other categories, such as partially proficient, are dependent on the primary goal of proficiency. Two methods of data collection were employed: survey and think-aloud procedures. Participation was voluntary.

Survey questionnaire. Participants completed a survey at the conclusion of each of the three rounds. The surveys varied somewhat from round to round to reflect each rounds' uniqueness. The surveys, which consisted of open- and closed-ended questions, addressed participants' thought processes as their judgments evolved from round to round. The closed-ended questions targeted participants' understanding of the Bookmark procedure, strategies used to select bookmark placements, and factors influencing bookmark placements (e.g., personal experiences with students, opinions expressed by other participants). The open-ended questions invited participants to include other considerations or elaborate on their responses to the closed-ended questions. Copies of

the surveys may be obtained from the first author. Usable survey data was obtained from 67 of the 69 participants.

Think-aloud procedure. Table leaders provided additional data for the study by participating in taped think-aloud procedures after each round. Table leaders were selected because they would provide representation from each small group and would likely be willing, given their participation in extra training. All three table leaders in SS1 and six of the eight table leaders in SS2 agreed to participate. However, it was not possible to obtain think-aloud activities from each of the nine table leaders after each round. Because small groups completed their judgments at different times, there was not sufficient time between the end of Round 2 and the beginning of Round 3 to complete the think-aloud procedure. Nine table leaders participated in Round 1, seven table leaders participated in Round 2, and the same seven participated in Round 3.

The think-aloud procedure required table leaders to describe their thoughts or thought processes (Ericsson & Simon, 1999) for deciding on the proficient bookmark placement. Table leaders independently summarized their thoughts using a hand-held tape recorder. At the completion of each round, table leaders were taken to a spacious room with ample separation of the table leaders to allow for non-interference and privacy. Instructions were provided orally, as well as typed on paper for reference. The instructions were:

Please use the tape recorder to simulate your thought processes you went through to place your proficient bookmark. You may refer to specific items, your understanding of what the bookmark means, and other considerations you thought of when making your decision.

Table leaders were permitted to use their item maps and ordered item booklets to complete the task.

The tapes of the think-aloud procedures were transcribed. Protocol analysis of the transcripts was conducted to identify cognitions; that is, information participants attended to when making their decisions (Ericsson & Simon, 1999). The first author, who has experience coding qualitative data, developed the coding scheme to reflect these cognitions or themes. Each theme was coded if present in a transcript, regardless of the extent to which the theme was elaborated. Appendix A contains definitions of the coding categories and examples. Inter-rater reliability was conducted on the entire set of transcripts. The second rater is a doctoral student in an educational measurement program. Percentage of agreement was 83.1%, indicating good reliability.

Results

Survey Results

Understanding of the Bookmark Procedure

Round 1. Participants were asked to rate their ease of understanding the definition of mastery. Refer to Table 1 for a summary of the results by standard setting. Over 80% of participants from each standard setting committee indicated the definition was very easy ($n_{SS1} = 17, 77\%$; $n_{SS2m} = 9, 39\%$; $n_{SS2s} = 12, 55\%$) or somewhat easy ($n_{SS1} = 4, 18\%$; $n_{SS2m} = 10, 13\%$; $n_{SS2s} = 10, 45\%$) to understand. Only one participant (5%) in SS1 endorsed that the definition was somewhat difficult to understand. On the SS2m committee, three participants (13%) indicated the definition was somewhat difficult to understand and one participant (4%) indicated that the definition was very difficult to understand.

To assess participants' understanding of the definition of mastery, the Round 1 survey included two questions. The first question related to students' chances of success on the set of items preceding the proficiency bookmark. The second question related to students' chances of success on the set of items following the proficiency bookmark. The two response options were: less than a 2/3 (67/100) chance of success, and equal to or greater than 2/3 (67/100) chance of success. Ninety percent or higher of the participants from each committee correctly answered question 1 ($n_{SS1} = 21, 91\%$; $n_{SS2m} = 22, 100\%$; $n_{SS2s} = 20, 91\%$) and question 2 ($n_{SS1} = 20, 90\%$; $n_{SS2m} = 22, 100\%$; $n_{SS2s} = 22, 100\%$).

Round 2. In Round 2, participants were asked to indicate whether their understanding of the Bookmark procedure had changed from Round 1. Over half of the participants in SS1 (64%) endorsed having the same understanding of what it means to place their bookmark in Round 2 as they did in Round 1, whereas over half of the participants in SS2m (65%) and SS2s (55%) indicated they had a better understanding in Round 2 compared to Round 1. No one indicated having a better understanding in Round 1 than in Round 2. Refer to Table 2 for a summary of the results. Round 2 data is presented on the first line; Round 3 data is presented on the second line in italics.

Round 3. Results indicate that over 80% of participants in SS1 (81%) and SS2m (86%), and approximately 60% of participants in SS2s (59%) had the same understanding of the Bookmark procedure in Round 3 as they did in Round 2. Less than 20 % of participants in SS1 (19%) and SS2m (14%) indicated they had a better understanding in Round 3 than they did in Round 2, whereas one-third of the participants in SS2s indicated such. Two participants in SS2s (9%) endorsed having a better understanding in Round 2 than in Round 3.

Item Selection Strategy

Round 1. Participants were asked to indicate the strategy they used to determine the placement of their Round 1 proficient bookmark. Table 3 contains the results of item selection strategies presented by round and standard setting committee. Round 1 results are presented in the first line of each row. In SS1, two item selection strategies were primarily used. Approximately half of the participants (52%) in SS1 indicated that they identified an interval of items and selected an item within the interval. The mean number of items in the interval was 5.50, with a range of 3 to 8 items ($SD = 2.17$). Approximately one in four participants (26%) indicated that they identified a single item and focused on the skills assessed by that item. Five participants (22%) provided written comments, but the comments were not codable as a strategy for item selection; rather, the comments pertained to factors of influence (e.g., content analysis, personal experience, difficulty level of items).

In SS2, three item selection strategies were primarily used. Slightly more than half of the participants in the math content area (55%) identified a single item and focused on that item and the preceding items. The next most frequent item selection strategy, endorsed by approximately one quarter of the respondents (27%), was to identify an interval of items and select an item within that interval. The mean number of items in the interval was 6.25, ranging from 4 to 10 items ($SD = 2.63$). Only one person indicated that he identified a single item and focused on the skills assessed by that item, while another person commented that he identified a single item and focused on that item and items that followed in the ordered item booklet. Two participants' comments (9%) were not codable as an item selection strategy.

In the science content area, slightly more than half (55%) of the participants identified an interval of items and selected an item within the interval. The mean number of items in the interval was 7.09 ($SD = 6.74$, with a range of 4 to 27 items). Other participants (41%) identified a single item and focused on that item and the preceding items. One person's comments did not pertain to an item selection strategy.

Round 2. The item selection strategies in Round 2 were limited almost exclusively to two response options. Refer to the second line of data presented in italics in Table 3. In SS1, more than half of the participants (64%) indicated they identified an interval of items and selected an item within that interval ($M = 5.00$, $SD = 2.83$, range of 3 to 12 items). Approximately one quarter of the respondents (23%) identified a single item and focused on that item and the preceding items. Three participants (14%) noted comments that were related to influences of their decisions rather than an item selection strategy.

In SS2m, participants' responses were equally divided ($n = 11$, 48%) between the options of identifying an interval of items ($M = 6.56$, $SD = 2.24$, with a range from 4 to 10 items) and identifying a single item and focusing on that item and the preceding items. One person noted a strategy that could not be coded as an item selection strategy.

In SS2s, the most frequently endorsed strategy (41%) was to identify a single item and focus on the skills assessed by that item and the preceding items. The next most frequently endorsed strategy (36%) was to identify an interval of items and select an item within the interval ($M = 5.43$, $SD = 2.15$, with a range of 3 to 12 items). In addition, two participants (9%) identified a single item and focused on the skills assessed by that item.

The comments of three people (14%) were not codable as an item selection strategy (e.g., group discussion, item content).

Round 3. Similar to the results of Round 2, the main strategies participants used for selecting their proficient bookmark location were: selecting an item from an interval of items, and identifying a single item and focusing on that item and the preceding items. Refer the third line of data presented in bold in Table 3. In SS1, the most frequently chosen strategy was selecting an interval (43%). The mean number of items in the interval was 3.40 ($SD = 0.55$, with a range of 3 to 4 items). Approximately one quarter of the respondents (26%) provided comments that were not codable as an item selection strategy. Five participants (22%) identified a single item and focused on that item and the preceding items, whereas one person identified a single item and focused on the skills assessed by that item. One participant indicated that he/she used another strategy, but did not specify what that strategy was.

In SS2m, the most frequently chosen item selection strategy was to identify an interval of items (43%). The mean number of items in the interval was 5.40 ($SD = 2.79$, with a range of 3 to 10). The next most frequently chosen strategy was to identify a single item and attend to that item and the preceding items (30%). Five participants (22%) provided comments that did not reflect an item selection strategy (e.g., no change from round 2; change based on group discussion; attended to content). One person identified a single item and concentrated on the skills assessed by that item only.

In SS2s, one-third of participants identified a single item and focused on that item and the preceding items. The next most frequently chosen strategy, selected by 29% of the participants, was to identify an interval of items ($M = 5.00$, $SD = 1.00$, with a range of

4 to 6 items). Six participants (29%) provided comments that did not reflect an item selection strategy (e.g., no change from round 2; attended to content, item difficulty, or failure rate). Two people (10%) identified a single item and focused on the skills assessed by that item.

Factors of Influence

Round 1. Participants were asked to identify the factors they considered when selecting their proficient bookmark placement. A list of possible influences was provided with instructions to check all that applied. In addition, participants were invited to include other influences that were not listed. The results for Round 1 are summarized in the first line of data presented in Table 4. In SS1, at least half of the participants indicated that they relied on the following influences: personal experiences working with students (91%), knowledge of the state content standards (78%), and their understanding of the performance level descriptors (57%). The remaining influences were endorsed less frequently: opinions expressed by small group members (30%), and opinions expressed by a single group member (4%). Other influences participants noted from open-ended responses were the difficulty level of items ($n = 2$, 9%), and coverage of the reporting categories ($n = 2$, 9%). These options were integrated into the surveys for SS2.

In SS2m, the following four influences garnered the support of at least half of the participants: difficulty level of items (91%), personal experiences working with students (82%), knowledge of state content standards (77%), and understanding of performance level descriptors (68%). The remaining options were endorsed by fewer than half of the participants. These included the opinions expressed by small group members (45%),

coverage of reporting categories (23%), and the opinions expressed by a single small group member (14%).

In SS2s, similar results were found. The same four influences were endorsed by more than half of the participants: personal experiences working with students (77%), understanding of performance level descriptors (64%), difficulty level of items (64%), and knowledge of state content standards (59%). Fewer than half of participants endorsed the remaining options: opinions expressed by small group members (36%), opinions expressed by a single small group member (5%), and coverage of reporting categories (5%).

Round 2. Participants who made changes to their proficiency bookmarks in Round 2 were asked to indicate the factors that influenced their decisions. The results for Round 2 are summarized in the second line of data presented in Table 4. Consistent across all three standard settings, participants endorsed that they were influenced by the opinions expressed in the small group discussion (100% for SS1, 95% for SS2m, 87% for SS2s). Only four people (20%), participants in SS2m, indicated that they were influenced by the opinions expressed by a single panel member. Participants also relied on their experiences working with students in SS1 (67%) and SS2m (55%), but did not refer to their experiences as readily SS2s (27%). The difficulty level of items, an option not included on the SS1 survey, was a consideration for 90% of the participants in SS2m and for 60% of participants in SS2s. In contrast to the findings in Round 1, only one person, a participant in SS2s, indicated his/her knowledge of state content standards was a consideration in the Round 2 judgment. Similarly, two people each from the SS1 (17%) and SS2m (10%) referred to their understanding of performance level descriptors.

Round 3. Participants who changed their proficiency bookmark in Round 3 were asked to indicate the influences in their decisions. Refer to the numbers in bold print in

Table 4. More than half of the participants in Round 3 indicated they were influenced by the Round 3 discussion to change their proficient bookmark (92% in SS1, 83% in SS2m, 63% in SS2s). A third of participants or less were influenced by the opinions of a single individual in Round 3 discussion (25% in SS1, 33% in SS2m, 19% in SS2s). Difficulty level of the items was frequently indicated as a consideration for participants in SS2 (73% in SS2m; 75% in SS2s). One-third of participants in SS1 and SS2m relied on their personal experiences working with students, whereas over half relied on these experiences in SS2s (56%).

Change of Proficient Bookmark Location.

Round 2. After Round 2 discussion, participants made a second placement of their bookmarks. A summary of the round to round changes to proficiency bookmark locations appears in Table 5. The first line of data refers to the Round 2 changes and the second line of data in italics refers to the Round 3 changes. In SS1, there were similar numbers of participants who maintained ($n = 8$, 38%), raised ($n = 7$, 33%), and lowered ($n = 6$, 29%) their proficient standard. Likewise, there were similar numbers of participants in SS2s who maintained ($n = 7$, 32%), raised ($n = 7$, 32%), and lowered ($n = 8$, 36%) their proficient standard. Slightly more than half of the SS2m participants ($n = 12$, 52%) raised their standard, while others lowered ($n = 8$, 35%), or maintained ($n = 3$, 13%) their standard.

Round 3. In SS1, slightly less than half of the participants ($n = 11$, 48%) maintained their proficiency bookmark location from Round 2, while others lowered ($n = 7$, 30%) or raised ($n = 5$, 22%) their standards. Over half of participants lowered their proficiency bookmark from Round 2 judgments in SS2 ($n_{SS2m} = 13$, 57%; $n_{SS2s} = 16$,

73%). For the remaining participants in SS2m, equal numbers of participants ($n = 5$, 22%) maintained and raised their bookmarks. For the remaining participants in SS2s, there were more that maintained ($n = 5$, 23%) than raised ($n = 1$, 5%) their proficient standard.

Think-Aloud Results

Round 1

Several distinct themes emerged from analysis of the nine table leaders' Round 1 think-aloud procedures. While setting their proficient bookmarks, table leaders primarily relied on their personal experiences working with students, the content of the test, and the consequential validity of finding an appropriate location for the bookmark. Table 6 contains the think aloud themes by standard setting and round. Results for Round 1 may be found on the first line of each row.

Personal experiences working with students. All table leaders relied on their experiences working with students when making their proficient judgments. Seven table leaders commented on their ability to identify the content students are capable of answering and/or anticipate the mistakes they would make. One table leader noted, "My main thing I probably used more than anything else was my personal experience with students. I do teach at a school where we have a very wide range of abilities from the very top to the very bottom." Because of her experience, she had insight into "what is easier for them and what is more difficult for them." Two table leaders visualized a particular type of student (i.e., average, just barely proficient) when determining their proficient bookmark.

Content. All table leaders indicated that they explored the content of the test before determining their proficient bookmark placement. Several perspectives on content emerged. One sought a balance between the content areas (“they separated evenly into statistics, geometry and algebra like they should have”); one looked for a place that reflected basic coverage of the content (“I felt like those were items that were basic understanding, basic knowledge.”); one attended to the number of steps it took to solve the problem (“That’s where the skill level changed. From the simple I can answer it fairly quickly, one or two steps, to where I have to use quite a few steps to complete the problem.”). Other considerations included attending to the abstractness of the items, the application of concepts, and the readability of items, attending to the wordiness and vocabulary.

Consequential validity of the bookmark. Six table leaders noted the importance of selecting an appropriate bookmark. One table leader commented on students’ opportunity to learn:

If all students were taught according to the information that was asked for in the standards, if all students were given that opportunity, I felt that where I placed my bookmark was at an appropriate place. All students would not necessarily know everything up until that point, but I think all students would have the opportunity to learn that information because the teachers would have had an opportunity to teach it.

Another table leader touched on students writing the exam in their second language:

I want to give every student a fighting chance.... When I look here, the stuff in front of the bookmark, I think regardless of language barrier or reading skills or

any type of bias, I think this part encompasses what all that kids can do. It is math for math's sake. If we have a student come over from Mexico, they could still look at the problem and say, yeah, I've done that before, even though the language is a little bit different, I've seen that kind of a problem and it is not confusing.

Other comments reflected consideration for setting realistic standards. One table leader noted that many students would not be attending college; another mentioned that students' futures would be affected because passing is a requirement for graduation; one referenced her concern for consequences to the education board, citing news stories from other states that reported the percent of students failing.

Small group discussion. One participant said he selected his bookmark at the location where there was group discussion in completing the item map.

Round 2

The themes from protocol analysis of the Round 1 think-aloud procedures were reiterated and new themes were identified in Round 2. Seven table leaders completed the think-aloud procedure. In order of most frequently mentioned, table leaders considered the content of the test, consequential validity of setting the bookmark, opinions expressed during small group discussion, opinions expressed by classroom teachers, and the results from Round 1. One person reflected on her personal experience working with students.

Round 2 results may be found in Table 6. Refer to the italicized print.

Content. Six table leaders re-evaluated their judgments based on the content of the test. For example, one changed her bookmark because she had overlooked one content area in her initial proficiency placement. Another realized there were more

cognitive skills necessary for some of the problems than she originally thought, thereby underestimating the difficulty of the items in Round 1. Another reassessed the items in light of reading difficulty and the number of word problems.

Consequential validity of the bookmark. Five table leaders alluded to setting realistic standards that are fair to the students. One table leader reiterated her Round 1 concern that the cut score will impact the lives of many of children. Two table leaders commented that passing the exam is required for graduation and many students may not require the skills in their work or further education. Another strived for a standard that was attainable by all students. One table leader commented that some teachers are not certified at the secondary level for the content area, and speculated that some content may not be covered in class.

Small group discussion. Round 2 discussions were instrumental in five table leaders' Round 2 judgments. Table leaders alluded to the discussion, most noting changes in their perceptions of the content covered in the exam.

Opinions of classroom teachers. Four table leaders commented that they listened carefully to the opinions expressed by teachers in their small group. One stated, "I felt like they had a good pulse as to what was going on and the skills that could be taught for kids."

Results of Round 1. Four table leaders referred to the results of their group's judgments from Round 1. Three table leaders lowered their proficiency bookmark to be more in keeping with the teachers ($n = 2$) or other group members ($n = 1$). Another table leader commented that there were several judgments above and below his placement, but he maintained his original placement, noting that he did not see reason to change.

Personal experience working with students. Unlike the results of Round 1 where everyone considered their experience working with students, only one table leader mentioned her experience working with students by anticipating students' chance of success on a group of items in her Round 2 deliberation.

Round 3

The themes that emerged in the first two rounds were reiterated in Round 3. In order of most frequently mentioned were: consequential validity of setting the bookmark, content, opinions expressed in large group discussion, results from the Round 2 judgments, opinions expressed by teachers, opinions expressed in small group discussion, and personal experience working with students. Round 3 results may be found in Table 6. Refer to the bold print.

Consequential validity of the bookmark. Six table leaders acknowledged the importance of setting the bookmark at a position that is fair to the students. A couple of the table leaders strived for a balance between setting a fair standard, yet maintaining high expectations for the students. One stated,

I wanted to make it high enough that it sets a good standard. I didn't want it too low. It would mean that we didn't have good expectations for our students...but again not so high that it makes it too challenging for a lot of our students.

Another table leader commented on her choice of proficient bookmark, "I felt like it was the best for students. I don't think we sacrificed any content, and I don't think we sacrificed students for it."

Content. Five table leaders reconsidered to the content of the test in Round 3. For example, one table leader noted that she lowered her bookmark after going over the

item map and deciding to exclude some of the content she previously thought was necessary for marginal students.

Large group discussion. Four table leaders reconsidered the content of the exam after listening to the items discussed in the large group. One commented, “After listening to the group talk and listening to all the reasons why it should be here or there, I really felt like [my placement] was a good place. I felt comfortable with my decision.”

Results of Round 2. Three people noted the results of Round 2. One person decided to match her bookmark to the median; another commented on the consistency of his placement with that of the median of the groups; one table leader lowered her bookmark after discovering the percent of students that would fail with the Round 2 cutscores.

Opinions of classroom teachers. Two table leaders commented that they were influenced by the contributions of the teachers.

Small group discussion. One table leader made reference to the diversity of student populations that were represented by her small group members.

Experience working with students. Like Round 2, only one table leader alluded to her experiences working with students. She commented on the kinds of students at her school and the content that they would correctly answer.

Discussion

The present study sought to investigate the cognitive experiences of Bookmark participants through the use of survey and think-aloud methods following each round of judgments. Results were consistent across the three standard setting committees, with

one exception. The findings will be discussed as they pertain to the three research questions.

Understanding of the Bookmark Procedure

The definition of mastery is introduced to participants prior to their Round 1 judgments. Findings from the Round 1 survey indicated that participants found the definition of mastery easy to understand and were able to interpret the definition correctly. The definition of mastery is central to the judgments participants make for each performance standard.

Slightly more participants indicated that their understanding of the Bookmark procedure improved from Round 1 to Round 2, but a clear majority had the same understanding of the procedure in Round 3 as in Round 2. Therefore, some participants' understanding evolved from Round 1 to Round 2, but remained constant in Round 3; others had the same understanding across the three rounds. These results suggest that participants understand the process and contribute to the procedural evidence that may be used to evaluate the appropriateness of the performance standards derived from the Bookmark standard setting.

Item Selection Strategy

Participants primarily used two strategies for placing their bookmarks: identifying an interval of items and then specifying a location within that interval; and identifying a single item with consideration to the items that preceded it in the ordered item booklet. For those identifying an interval of items, the size of the interval decreased with each round, indicating less indecision. The finding is consistent with the instructions relayed in training and contributes to the procedural evidence supporting the interpretations of

standards set using the Bookmark procedure. Participants are instructed to conceptualize the bookmark placement as dividing the ordered item booklet into two sets of items, those that students at the border of a performance category should master and those that need not be mastered (Mitzel et al., 2001). In addition, on average, only 6% of the participants across the rounds indicated that they focused on a single item to make their decision, which refutes the criticism of the Bookmark procedure that participants are making their decisions based on a few items or a single item.

Factors Influencing Bookmark Selection

Participants changed the emphasis of influences they considered when placing their proficient bookmarks in each round. The results will be discussed in light of recent work that incorporates cognitive psychology concepts to understand the process of the judgment tasks participants perform.

Jaeger (1995) couched the judgment and decision tasks required of participants at standard settings into the framework of cognitive psychology. Jaeger utilized Johnson-Laird's theory of mental models (1981), which stipulates that people develop a representation of a problem and link other knowledge to the problem. Jaeger proposes that to build a mental model, uncertainties created by lack of completeness of the problem must be addressed. Although standard setting tasks encompass a number of uncertainties, uncertainties related to following the procedure include: participants' lack experience with the tasks requested of them (e.g., estimating the probability that a particular group of students will respond correctly to an item), and participants' ability to form mental images of abstract principles, such as a "minimally competent" or "marginal" student. To cope with uncertainty in a decision task, people retrieve

information from memory in an attempt to complete the representation of the problem (Berkeley & Humphreys, 1982).

Jaeger (1995) cites the information processing model of Anderson (1970) to understand the processes of the participants. Within this framework, the concepts of anchoring and adjustment offer insight into the sequence of mental operations panelists use to fashion their responses. An anchor point serves as an initial response to the problem.

Results of Round 1 reveal the influences participants relied on to create their initial anchors. Participants primarily referenced their experiences working with students, their knowledge of the state content standards, their understanding the performance level descriptors, and the difficulty of the items. The think-aloud activities also highlighted table leaders' reliance on their experiences with students. Experience often was mentioned in conjunction with identifying the content that students are capable of answering. Table leaders' expressed concerns for setting a standard that is fair and realistic for the students.

Round 1 results mirror what would be expected and desired. Participants are recruited based on their reputation in the education field. They typically have expertise in instruction, knowledge of the state content standards, and years of experience working with students. In the first round, participants make their judgments independently. They are discouraged from discussing the items in terms of what content should be mastered, so that the first round judgments are untainted by the opinions of others (Mitzel et al., 2001). Participants develop their initial anchors from that which they bring to the standard setting; that is, their own experiences and expertise.

The initial anchor then is adjusted in Rounds 2 and 3. Additional attributes of the problem are encoded to develop a revised representation of the problem. Jaeger states that when the process of mental modeling is “guided through appropriate decision aids, results should be more consistent and more likely grounded in appropriate bases for resolution of uncertainty” (1995, p. 59).

Jaeger (1995) proposed three recommendations to be incorporated into standard setting procedures. The first recommendation is sequential provision of information, which includes the use of normative data on students’ performance and consequences of the recommended standards. The second recommendation is guided social interaction that permits judges to evaluate their initial anchors by comparison with others, and to adjust accordingly. Examples include feedback on the judgments of other panelists and the opportunity to modify their initial judgments. The third recommendation is the introduction of alternative cognitive schema. Panelists may revise their mental models as they become aware of the experiences and opinions of others. Jaeger proposes that panelists formally present their rationales for their judgments.

These recommendations are features of the Bookmark procedure, as well as other standard setting procedures. Jaeger’s first recommendation involves the presentation of normative data on student performance. Participants receive normative information in the form of an ordered item booklet, where the items are ordered from easiest to hardest. This method of presenting normative data is described as useful because it helps avoid erroneous estimations of probability (Zieky, 2001). The decision to include impact data, however, is a decision made by the state department of education. Impact data is often presented during a Bookmark standard setting, as it is with other procedures.

The second recommendation, guided social interaction to permit judges to evaluate their initial anchors, is a feature of some standard setting procedures, including the Bookmark procedure. In Round 2, participants place tabs on pages corresponding to each group members' initial bookmark locations in the ordered item booklet. Participants learn the extent to which other group members' judgments are consistent with their own. In addition, participants are given the opportunity to adjust their initial bookmark placements in Rounds 2 and 3.

Recommendation three, the introduction of alternative cognitive schema, is a feature of the Bookmark procedure, as well as other standard setting procedures. Group discourse in Round 2 focuses on the items between the first and last bookmark placement for a given performance category. The table leader facilitates discussion of items in terms of skill level and academic content that should be mastered by students at a given performance level. Each participant explains his/her rationales for the bookmark placement. Results indicate that participants listened to the rationales of their peers; they cited the influences of small and large group discussions for their Rounds 2 and 3 judgments. Further, the composition of the small groups reflects the diversity of educational professionals across the state. Participants dialogue with other educational professionals who have a variety of experiences and opinions. After having the opportunity to compare their initial anchors with other group members and adjust their mental representations, participants make second judgments of their bookmark locations.

The Round 2 survey results reveal that participants relied extensively on the opinions expressed by their peers to make their judgments. Participants are expected to set standards for students across the state, not just for their own classrooms. One would

hope that the diversity of opinions represented by their peers would provide the insight necessary to do so, as is reflected in the data. The majority on each committee endorsed being influenced by the group discussion. That is, they were influenced by the variety of opinions expressed, rather than the opinions of a single group member. Discussion directed toward understanding item content resulted in the majority of participants reconsidering their bookmark placements in light of the difficulty level of the items. In contrast to Round 1 results, participants' personal experiences with students played a less prominent role in determining subsequent judgments.

The results of the think-aloud procedures also reflected the influence of group members' opinions. Table leaders particularly were attentive to the insights provided by teachers of the group. They also noted attending to the content of the exam when selecting their proficiency placements. Protocol analysis of the think-aloud procedures indicated table leaders deliberated over setting a valid bookmark and the previous round's results before placing their second proficiency bookmarks.

Like Round 2, the large group discussion in Round 3 focuses on the necessary skills and content participants infer from the items. Discussion revolves around the items between the lowest and highest group median locations. Results for Round 3 are similar to the results of Round 2. The survey findings indicate participants were influenced by the opinions voiced in Round 3 discussion, although several participants in each committee considered the opinions of a single panelist. Participants also reconsidered the difficulty level of the items and several panelists on each committee reflected on their experiences working with students. The think-aloud activities revealed participants' consideration of the group discussion. The consequences of setting the bookmark and the

content of the exam were themes that also emerged. Round 3 discussion provides participants the opportunity to further their understanding of others' experiences and refine their representations before casting final judgments.

There was consistency in findings across the two methodologies. The survey provided participants with a list of possible influences they may have considered in placing their proficient bookmark; however, the list was not exhaustive. The think-aloud procedure provided table leaders the opportunity to freely relay their thought processes. Both methods pointed to reliance on experiences with students in Round 1 and reliance on the opinions of their peers Rounds 2 and 3. Several additional themes were derived from the think-aloud protocol analysis: the implications of selecting a cut score on an exam required for high school graduation; and the results from the previous round entered into the decision-making process.

The only salient difference between the standard setting committees was in the number of participants maintaining and changing their decisions in Round 2 and 3. In the first standard setting, there were almost equal numbers of participants who maintained, raised, and lowered their proficiency bookmarks in Round 2. Almost half of the participants maintained their proficiency bookmark in Round 3, indicating certainty in their decisions. In the second standard setting, the majority of participants lowered their bookmarks in the third and final round, likely due to the inclusion of impact data.

In summary, the purpose of the study was to document the cognitive experiences of Bookmark participants for two reasons. The first reason was to provide procedural evidence supporting the interpretations of the standards set using the Bookmark procedure. Procedural evidence is important to evaluate the validity of the performance

standards, since few empirical checks on the performance standards are available (Kane, 2001). The second reason was to investigate the unique information provided by a Bookmark standard setting. The most common finding in the standard setting literature is that different standard setting procedures yield different results. It has been suggested that the answers may lie in investigating the standard setting procedures in terms of what is being judged, the tasks prescribed for participants, and the cognitive processes they use to formulate their decisions (Mitzel et al., 2001). However, few studies have sought to identify the specific information yielded by participants' judgments. The present study provides a measure of such information for the Bookmark procedure.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review*, *11*, 153-170.

Berkeley, D., & Humphreys, P. (1982). Structuring decision problems and the bias heuristic. *Acta Psychologica*, *50*, 201-252.

Ericsson, K. A., & Simon, H. A. (1999). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

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

Jaeger, R. M. (1995). On the cognitive construction of standard-setting judgments: The case of configural scoring. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments, Vol. II*, pp. 57-73. Washington, D.C.: U.S. Government Printing Office.

Johnson-Laird, P. N. (1981). Mental models in cognitive science. In D. A. Norman (Ed.), *Perspectives on cognitive science* (pp. 147-191). Hillsdale, NJ: Erlbaum.

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.

Lewis, D. M. (2001, April). *Standard setting challenges to state assessments: Synthesis, consistency, balance, comparability*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, Washington.

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

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Mitzel, H. C., Lewis, D. M., Patz, R. J., Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Erlbaum.

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. 19-88). Mahwah, NJ: Erlbaum.

Table I

Ease of Understanding Mastery Definition by Standard Setting Committee for Round 1

	SS1	SS2	
	Math (<i>n</i> = 22)	Math (<i>n</i> = 23)	Science (<i>n</i> = 22)
Very difficult	0	1 (4%)	0
Somewhat difficult	1 (5%)	3 (13%)	0
Somewhat easy	4 (18%)	10 (43%)	10 (45%)
Very Easy	17 (77%)	9 (39%)	12 (55%)

Table II

Understanding of the Bookmark Procedure in Rounds 2 and 3 by Standard Setting Committee

	SS1	SS2	
	Math	Math	Science
Understanding	(<i>n</i> = 22) ^a	(<i>n</i> = 23) ^b	(<i>n</i> = 22) ^b
Same understanding in this round as previous round	14 (64%) <i>17 (81%)</i>	8 (35%) <i>19 (86%)</i>	10 (45%) <i>13 (59%)</i>
Better understanding in this round than previous round	8 (36%) <i>4 (19%)</i>	15 (65%) <i>3 (14%)</i>	12 (55%) <i>7 (32%)</i>
Better understanding in previous round than this round	0 <i>0</i>	0 <i>0</i>	0 <i>2 (9%)</i>

Note. Round 2 data is presented in the first row; Round 3 data is presented in the second row in italics.

^aSample size for Round 3 is 21.

^bSample size for Round 3 is 22.

Table III

Item Selection Strategies for Location of Proficient Bookmark by Standard Setting Committee and Round

Item Selection Strategies	SS1	SS2	
	Math (<i>n</i> = 23) ^a	Math (<i>n</i> = 22) ^b	Science (<i>n</i> = 22) ^c
Identified an interval of items and selected an item within the interval	12 (52%) <i>14 (64%)</i> 10 (43%)	6 (27%) <i>11 (48%)</i> 10 (43%)	12 (55%) <i>8 (36%)</i> 6 (29%)
Identified a single item and focused on that item and the preceding items ^d	– <i>5 (23%)</i> 5 (22%)	12 (55%) <i>11 (48%)</i> 7 (30%)	9 (41%) <i>9 (41%)</i> 7 (33%)
Identified a single item and focused on the skills assessed by that item	6 (26%) <i>0</i> 1 (4%)	1 (5%) <i>0</i> 1 (4%)	0 <i>2 (9%)</i> 2 (10%)
Other, but comment is not strategy related	5 (22%) <i>3 (14%)</i> 6 (26%)	2 (9%) <i>1 (4%)</i> 5 (22%)	1 (5%) <i>3 (14%)</i> 6 (29%)
Other strategy	0 <i>0</i> 1 (4%)	1 (5%) <i>0</i> 0	0 <i>0</i> 0

Note. Round 1 data is presented in the first row; Round 2 data is presented in the second row in italics;

Round 3 data is presented in the third row in bold.

^aSample sizes for Rounds 2 and 3 are 22 and 23 respectively.

^bSample size for Rounds 2 and 3 is 23.

^cSample sizes for Round 2 and 3 are 22 and 21 respectively.

^dOption was not included in the Round 1 survey for SS1.

Table IV
*Factors Participants Endorsed as Influencing the Proficient Bookmark Location by
 Standard Setting Committee and Round*

Influence	SS1	SS2	
	Math (<i>n</i> = 23) ^a	Math (<i>n</i> = 22) ^b	Science (<i>n</i> = 22) ^c
Personal experiences with students	21 (91%) 8 (67%) 4 (33%)	18 (82%) 11 (55%) 6 (33%)	17 (77%) 4 (27%) 9 (56%)
Knowledge of state content standards	18 (78%) 0 0	17 (77%) 0 2 (11%)	13 (59%) 1 (7%) 0
Understanding of performance level descriptors	13 (57%) 2 (17%) 1 (8%)	15 (68%) 2 (10%) 4 (22%)	14 (64%) 0 2 (13%)
Difficulty level of items	2 (9%) ^d - -	20 (91%) 18 (90%) 13 (72%)	14 (64%) 9 (60%) 12 (75%)
Opinions expressed by small group members	7 (30%) 12 (100%) 1 (8%)	10 (45%) 19 (95%) 8 (44%)	8 (36%) 13 (87%) 3 (19%)

Table IV (continued)

	SS1	SS2	
Influence	Math (<i>n</i> = 23) ^a	Math (<i>n</i> = 22) ^b	Science (<i>n</i> = 22) ^c
Coverage of reporting categories	2 (9%) ^d	5 (23%)	1 (5%)
	-	5 (25%)	5 (33%)
	-	7 (39%)	3 (19%)
Opinions expressed by a single small group member	1 (4%)	3 (14%)	1 (5%)
	<i>0</i>	<i>4 (20%)</i>	<i>0</i>
	0	1 (6%)	0
Opinions expressed in Round 3	-	-	-
	-	-	-
	11 (92%)	15 (83%)	10 (63%)
Opinions expressed by an individual in Round 3	-	-	-
	-	-	-
	3 (25%)	6 (33%)	3 (19%)

Note. Round 1 data is presented in the first row; Round 2 data is presented in the second row in italics;

Round 3 data is presented in the third row in bold.

^aSample size for Rounds 2 and 3 is 12.

^bSample sizes for Rounds 2 and 3 are 20 and 18 respectively.

^cSample sizes for Rounds 2 and 3 are 15 and 16 respectively.

^dOption was not included in the survey for SS1, but was coded from open-ended responses.

Table V

Round to Round Changes to Proficient Bookmark Locations by Standard Setting Committee

Change	SS1	SS2	
	Math (<i>n</i> = 21) ^a	Math (<i>n</i> = 23) ^a	Science (<i>n</i> = 22) ^b
Maintained standard	8 (38%)	3 (13%)	7 (32%)
	<i>11 (48%)</i>	<i>5 (22%)</i>	<i>5 (23%)</i>
Raised standard	7 (33%)	12 (52%)	7 (32%)
	<i>5 (22%)</i>	<i>5 (22%)</i>	<i>1 (5%)</i>
Lowered standard	6 (29%)	8 (35%)	8 (36%)
	<i>7 (30%)</i>	<i>13 (57%)</i>	<i>16 (73%)</i>

Note. Round 1 data is presented in the first row and Round 2 data is presented in the second row in italics.

^aSample size for Round 3 was 23.

^bSample size for Round 3 was 22.

Table VI

Think Aloud Themes by Standard Setting Committee and Round

Theme	SS1	SS2	
	Math (<i>n</i> = 3)	Math (<i>n</i> = 4) ^a	Science (<i>n</i> = 2)
Content	3 (100%)	4 (100%)	2 (100%)
	3 (100%)	2 (100%)	1 (50%)
	1 (33%)	2 (100%)	2 (100%)
Personal experiences working with students	3 (100%)	4 (100%)	2 (100%)
	0	1 (50%)	0
	0	1 (50%)	0
Consequential validity of the bookmark	2 (67%)	2 (50%)	2 (100%)
	1 (33%)	2 (100%)	2 (100%)
	2 (67%)	2 (100%)	2 (100%)
Small group discussion	1 (33%)	0	0
	3 (100%)	1 (50%)	1 (50%)
	0	1 (50%)	0
Opinions of classroom teachers	–	–	–
	1 (33%)	2 (100%)	1 (50%)
	1 (33%)	1 (50%)	0

Table VI (continued)

Theme	SS1	SS2	
	Math (<i>n</i> = 3)	Math (<i>n</i> = 4) ^a	Science (<i>n</i> = 2)
Results from previous round	–	–	–
	<i>2 (67%)</i>	<i>1 (50%)</i>	<i>1 (50%)</i>
	2 (67%)	0	1 (50%)
Large group discussion	–	–	–
	–	–	–
	2 (67%)	1 (50%)	1 (50%)

Note. Round 1 data is presented in the first row; Round 2 data is presented in the second row in italics;

Round 3 data is presented in the third row in bold.

^aSample size for Rounds 2 and 3 is 2.

Appendix A

Think Aloud Coding Categories

Category Label	Description	Examples
Personal experience working with students	Table leader relies on his/her experiences working with students when making the judgment.	Table leader is able to identify what students find easy or difficult, or what students are capable of doing. Table leader is able to conceptualize a particular type of student (average, just barely passing).
Content	Table leader considers the content of the test when making the judgment.	Table leader indicates that he/she considered the content or makes reference to a specific aspect of the content (eg., content coverage, specific content such as geometry, number of steps required to solve the problem, readability of the items, item difficulty).
Consequential validity of placing the Bookmark	Table leader considers the importance and/or consequences of placing the bookmark when making	Table leader acknowledges the importance of selecting the bookmark by noting the consequences of the cutscore on students (eg., students are required to pass for high school graduation) or

	<p>the judgment.</p>	<p>the education board (eg., media stories from other states reporting the percentage of students who fail). Table leader makes reference to the importance of selecting an appropriate bookmark location by considering fairness to students, including students' opportunity to learn (eg., quality of teaching, exposure to material covered in the standards), or the population of students writing the exam (eg., sociodemographic background of the students, students with limited reading skills or English proficiency). Table leader strives for a reasonable standard (eg., selecting a standard that most students would be able to achieve, maintaining high expectations that are not too challenging).</p>
<p>Small group discussion (Rounds 1 and 2)</p>	<p>Table leader makes reference to the opinions expressed by other participants in small group discussion when making the</p>	<p>When the small group discussion influenced a table leader to reconsider his/her placement with regard to content coverage, group results from the previous round, or the implications of the cut score, code this category and other relevant category.</p>

	judgment.	
Opinions of classroom teachers	Table leader makes reference to the opinions of classroom teachers expressed in discussion.	
Results of previous round (Rounds 2 and 3)	Table leader makes reference to the results of group judgments from the previous round.	In Round 2, results from Round 1 are presented to the table leaders in the forms of median placement of the group and each member's placement. In Round 3, the median group scores from Round 2 are presented to the large group. In addition, impact data were provided at SS2.
Large group discussion (Round 3 only)	Table leader indicates that he/she considered the opinions expressed in the large group discussion.	