Combining the Best of Two Standard Setting Methods: The Ordered Item Booklet Angoff

Russell W. Smith*, Susan L. Davis-Becker and Lisa S. O’Leary
Alpine Testing Solutions, 51 West Center Street, #514, Orem, UT 84057; Russell.Smith@alpinetesting.com, Susan.DavisBecker@alpinetesting.com, Lisa.OLeary@alpinetesting.com

Abstract

This article describes a hybrid standard setting method that combines characteristics of the Angoff (1971) and Bookmark (Mitzel, Lewis, Patz & Green, 2001) methods. The proposed approach utilizes strengths of each method while addressing weaknesses. An ordered item booklet, with items sorted based on item difficulty, is used in combination with individual item ratings rather than a bookmark placement. The ordered item booklet attempts to make individual item ratings more palatable by providing some amount of information about item difficulty to the Subject Matter Experts (SMEs) participating in the standard setting. Allowing for individual item ratings attempts to address the issue of SMEs disagreeing with the order of items in the booklet in a Bookmark standard setting. This article illustrates the method as it was applied to a healthcare examination used for state licensure, comparing it with a more traditional Angoff approach. Results suggest that the proposed method has promise for future use pending further exploration.

Keywords: Angoff Method, Bookmark Method, Standard Setting

1. Introduction

There are many ways to recommend passing standards for criterion-referenced measures of knowledge, skills, and abilities. Typically, these methodologies are classified as test-based or examinee-based (Jaeger, 1989). Test-based methods focus on Subject Matter Experts (SMEs) serving as panelists and making judgments about expected examinee performance on the test content itself. Examinee-based methods require an external criterion variable to classify examinees or SMEs to make explicit judgments about knowledge or proficiencies of examinees. Examinee performance is then used to estimate the recommended cut score. To recommend passing standards for many high-stakes examinations, practitioners often utilize test-based methods when panelists lack familiarity with the skills and abilities of individual examinees or when insufficient numbers of examinees are available to use examinee-based methods. Judgments collected using test-based methods may be perceived as more objective than examinee-based approaches. Hambleton and Pitoniak (2006) described numerous test-based methodologies that have been documented in the literature and used in practice (e.g., Angoff, Bookmark, Ebel, Nedelsky, Jaeger-Mills). Two methods have emerged as prevalent in standard setting literature: Angoff (1971) and Bookmark (Mitzel, Lewis, Patz & Green, 2001).

Despite the further development of standard setting methodologies in recent years, some experts in the measurement community still express concerns regarding the reliability of the resulting cut scores; the comparability of cut scores obtained from differing standard setting approaches; and (c) panelist understanding of the methodology, process, and confidence in their results (Cizek & Bunch, 2007; Jaeger, 1989; Peterson, Schulz & Engelhard, 2011). Specifically, Cizek and Bunch (2007) indicated that due to the role of judgment in standard setting, a “knowable” parameter (i.e., a recommended cut score) could...
not be discovered (p. 19). Therefore, Peterson, Schulz, and Engelhard (2011) have suggested the utility of evaluating innovative standard setting methodologies against well-established ones. They noted that the “convergence of results of using two different standard setting methods has been cited as evidence of validity” (p. 5).

In this article, the authors describe a proposed hybrid standard setting method that combines the perceived advantages of the Angoff and Bookmark methods while attempting to mitigate the concerns about each method. The authors present a theoretical justification for the use of this hybrid method along with the comparative results of an implementation of this method versus an Angoff approach with a licensure testing program.

2. The Angoff and Bookmark Methods

The Angoff (1971) method requires panelists to evaluate each test item and estimate how a typical target examinee or group (e.g., minimally competent candidate, just proficient student) will perform on that item. There are two common variations of this approach. In the Yes/No variation of Angoff’s original methodology, panelists must make dichotomous decisions regarding whether or not the target examinee would answer the item correctly (Impara & Plake, 1997). In the more widely used Modified Angoff variation (described by Angoff (1971) in a footnote), panelists estimate the probability that the target examinee would answer the item correctly.

The more common variation of Angoff’s method is defended in the measurement community as more psychometrically defensible than the approaches of Nedelsky or Jaeger (Peterson et al., 2011). Berk (1986) compared the Angoff method to other judgment methods before the Bookmark method was established. Berk stated that “[T]he Angoff method appears to offer the best balance between technical adequacy and practicability” (p. 147). However, some members of the educational measurement and policy communities have criticized the Angoff method. Many members of the measurement community question the ability of panelists to make the necessary judgments about individual test items (e.g., Pellegrino, Jones & Mitchell, 1999; Shepard, Glaser, Linn & Bohrnstedt, 1993). For example, the Modified Angoff variation has been characterized as too cognitively complex for panelists given the number and specificity of judgments required while maintaining an appropriate interpretation of the key knowledge, skills, and abilities being tested (Pellegrino et al., 1999; Peterson et al., 2011; Shepard et al., 1993). In contrast, the Yes/No variation of the Angoff method is considered to be less cognitively complex than estimating probabilities of target examinees’ performance (Hambleton & Pitoniak, 2006). However, some criticisms point out that even this simpler method requires panelists’ judgments to evaluate the empirical difficulty of each item rather than an overall judgment about a set of items.

The Bookmark method also requires panelists to make judgments about the relative difficulty of the content represented by items on the test and estimate how a typical target examinee or group of such examinees will perform. Items are presented in an Ordered Item Booklet (OIB) based on previous item analysis or calibration results. Typically, one item or score point for constructed response items is presented on each page; the easiest item is on the first page of the booklet, and the most difficult item is on the last page. Panelists are presented with the OIB and asked to review each item in order (starting with the easiest item) and determine the position in the booklet where the target examinee would have a specified response probability of answering the item correctly. For example, if the specified response probability is 0.67, panelists place a bookmark dividing the booklet into two sections (when only one performance standard cut score is required) at the position in the booklet that they feel the target examinee no longer has at least a 67% chance of answering the items beyond that position correctly. The panelists should then continue to review the next few items and ensure they are beyond the expected capabilities of the target examinee or group of examinees. If they are, the bookmark should be left in place. If not, the panelists can move the bookmark until they reach the appropriate booklet position.

Arguably, the primary advantages of the Bookmark standard setting method are the reduced cognitive complexity of judging items through the use of an OIB and making a single judgment for each cut score rather than judgments about individual items. Panelists can use the ordering of the OIB to determine the relative empirical difficulty of the item and use this information in making their judgments. Although the Bookmark method requires fewer judgments than the Angoff method, it is still not without critics. Karatonis and Sireci (2006) note that the complexity of the Bookmark method and its task for panelists is still being questioned; comparisons to the
Angoff method have proven difficult given the response criterion, and one cannot be sure if panelists comprehend the task they are undertaking. These researchers continue on to point out a potential bias to Bookmark standard settings producing lower-than-intended standards given the difficulty in selecting the most appropriate response probability value for ordering the items presented to the panelists. Additionally, the Bookmark method has been criticized because panelists often disagree with item ordering by stating more difficult items appear to precede easier items and vice versa (e.g., Davis-Becker, Buckendahl & Gerrow, 2011; Karantonis & Sireci, 2006; Mitzel et al., 2001; Skaggs & Tessema, 2001). In one application of the Bookmark method, Skaggs and Tessema (2001) found that SMEs who questioned the booklet item ordering had much higher bookmark placements (i.e., recommended cut scores) than other panelists. In contrast, Davis-Becker et al. (2011) found that even when items were ordered randomly within a booklet, the average bookmark location was the same as with a properly OIB calling into question the criteria that panelists were using to make judgments. Finally, Bishop, Nhouyvanisvong, Zhu and Liassou (2011) described the challenge in determining a precise recommended theta cut score when large gaps exist between item difficulties, a problem that is more likely to exist on assessments with relatively fewer items.

When panelists perceive the empirically estimated item difficulty to be in contrast to conceptual difficulty, they are likely challenged in finding the appropriate location for the bookmark. For example, if panelists fear setting the cut score too high, they may place their bookmark early in the book (i.e., as soon as they find a single item they feel the target candidate would likely not answer correctly). Conversely, if panelists are worried about setting their cut score too low, they may place their bookmark further along in the book at the point where they are sure the target candidate would not likely answer any of the remaining items correctly.

3. Proposed Method

In the context of this ongoing debate about the merits of these two standard setting methods, the authors of this article investigated merging characteristics of these standard setting methodologies into a hybrid approach that capitalizes on the best features of each method and addresses some of the major criticisms of each method. This variation, the Ordered Item Booklet Angoff (OIB Angoff), presents examinees with the same type of ordered item booklet as would be used in a Bookmark standard setting. However, it allows panelists to predict performance by the target examinee for each item as part of the Angoff method (using Yes/No predictions or probability-based predictions).

In an application of the method, items within a test form are organized in an OIB so that the first item presented is the easiest and the last item is the most difficult. Panelists are informed of the order of the OIB and then asked to review each item and make a Yes/No Angoff-style judgment (which could also be an estimated probability or percentage) as to whether they believe the target candidate would likely answer the item correctly. Therefore, the design of this methodology incorporates the organization of the OIB from the Bookmark method with the judgmental task of the Angoff method. Although the design of the method encourages consideration for empirical item difficulty (by incorporating the OIB), panelists are allowed to provide whatever judgment they would like and thus disagree with the order of the items. Table 1 summarizes how the design of the OIB Angoff method addresses some of the primary criticisms leveled at the Angoff and Bookmark methods presented in previous sections.

Table 1 summarizes how the design of the OIB Angoff method addresses some of the primary criticisms leveled at the Angoff and Bookmark methods presented in previous sections.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angoff</td>
<td>Presents examinees with the same type of ordered item booklet as would be used in a Bookmark standard setting. Allows panelists to predict performance by the target examinee for each item as part of the Angoff method (using Yes/No predictions or probability-based predictions).</td>
</tr>
<tr>
<td>Bookmark</td>
<td>Items are organized in a manner that encourages panelists to predict performance by the target examinee for each item. Encourages panelists to provide whatever judgment they would like and thus disagree with the order of the items.</td>
</tr>
</tbody>
</table>

It is expected that panelists participating in the OIB Angoff method would make more “yes” judgments at the beginning of the booklets (i.e., easier items) and more “no” judgments at the end of the booklet (i.e., more difficult items). Likewise, it is expected that these panelists would estimate higher probabilities for items at the beginning of the booklet and lower probabilities for items at the end of the booklet. Thus, this approach increases the flexibility of the Bookmark method by allowing panelists to deviate from the expected pattern and disagree with the item booklet ordering scheme. The extent to which panelists deviate from the Guttman-like pattern (i.e., all 1’s followed by all 0’s) or higher probabilities at the beginning followed by lower probabilities at the end may indicate their level of disagreement with the empirical difficulty of the items.

To empirically investigate the feasibility of this approach, a study was conducted whereby a standard setting panel for a regional licensure examination was split so that half the panel followed the Yes/No Angoff methodology and the other half of the panel followed the OIB Angoff methodology.
4. Methods

4.1 Examination and Panelists

The examination used in this study was developed by a healthcare licensure testing agency that serves a consortium of state licensing boards. The examination is designed to be used by states as one requirement for state licensure eligibility. Each form of the examination consists of 50 dichotomously scored, single-response, multiple-choice items.

A single test form was used for the standard setting workshop. This form was one of two new pre-equated test forms that were created for this particular licensure exam. The test forms were pre-equated based on a pilot test with approximately 100 candidates across four different administrations. This sample was deemed sufficient for the purposes of creating these test forms and as a basis for this standard setting because of the representativeness in comparison to the full population.

Most implementations of the Bookmark method use Item Response Theory (IRT) parameters as a basis for creating the OIB and determining the cut score. Given the need in the OIB Angoff method to translate Angoff judgments (Yes/No, sum of “yes”) into a recommended cut score, the estimated item-level p-values were used to create the OIB in this study. IRT was used to estimate what the item p-values would be if the items had been administered to all examinees in the sample. This approach is in line with past work by Buckendahl, Smith, Impara, and Plake (2002) that attempted to simplify the Bookmark methodology by creating an OIB using item p-values and using the average page number of the bookmark placements as the recommended cut score because page numbers also corresponded to total score in that application. Although Cizek (2007) criticized this approach as “an incorrect implementation of the Bookmark method”, he went on to say that “this alternative… does provide a simple strategy and a reasonable alternative provided that all related training, materials, feedback, and so on are similarly realigned” (p. 90).

The 13 panelists that participated in this study consisted of educators (n=4), recently licensed practitioners (n=3), regulatory practitioners (n=3), and more experienced practitioners (n=3). Panelists were selected to be representative of the states served by the regional testing association.

4.2 Training

The workshop began with a training provided to the full panel (13 panelists). This training included an overview of the workshop and the purpose of standard setting, along with explanations of both methodologies (Angoff and OIB Angoff), the sample of candidates on which the item analysis was conducted, how the empirical item difficulties (i.e., p-values) were estimated, and how the ordering of the items in the OIB was determined.

A preliminary definition of the target candidate, referred to as the “Minimally Competent Candidate” (MCC), was then presented to the panelists. The MCC was described as having the minimum knowledge and skills necessary for licensure. A discussion was then facilitated with all 13 panelists as a single group, during which

<table>
<thead>
<tr>
<th>Table 1. Criticisms addressed by OIB Angoff Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Angoff</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Bookmark</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Combining the Best of Two Standard Setting Methods: The Ordered Item Booklet Angoff

the panelists discussed and documented the knowledge, skills, and abilities of that MCC. Notes regarding this discussion were taken on a flip chart, transcribed, and provided as a handout to be used by the panelists for the remainder of the workshop.

4.3 Operational Standard Setting

The panelists were then divided into two groups using a stratified random sampling approach where the strata were profession (educator or practitioner), years of experience, and geographical location within the region. From this point forward, the two groups were treated as independent standard setting panels. Each group was presented with a booklet containing the 50 items from a single form of the exam. For the Angoff group (7 panelists), the items were grouped by content area in no particular order within those content areas. For the OIB Angoff group (6 panelists), the items were presented in order of empirical item difficulty from the easiest item on the first page to the most difficult item on the last page. The Angoff rating process was reviewed and summarized to both groups with instructions to judge whether they believed the MCC or a group of MCCs would answer each item correct or incorrect (using the Yes/No variation). It was explained to the OIB Angoff group that the items in their booklets were ordered from easiest to most difficult based on the performance of the pilot sample candidates. No keyed responses or item statistics were provided to either group during Round 1.

After Round 1 ratings were completed by each of the panelists, feedback data based on the Round 1 ratings were provided within each group. No information was shared between groups. The feedback included individual recommended cut scores, the group mean and median recommended cut score, the range of recommended cut scores (plus and minus two standard errors of the mean), and the estimated pass rates associated with each possible cut score. Panelists were also given the correct answers and estimated p-values for each item for their reference during the second round of ratings.

To facilitate comparison of the results between groups, the study was designed so that the only intended difference between panels would be the order of the items. However, the Angoff group participated in a discussion between rounds regarding the alignment of the items to entry level practice, whether the content was addressed in training programs, and redundancy with a national board exam. While this discussion is viewed as tangential to the standard setting task and likely did not affect the item judgments, it may limit the comparability of the evaluation results between panels. Therefore, although evaluation results will be provided, the discussion in this article will primarily focus on the item ratings.

4.4 Evaluation

After Round 2 was complete and panelists provided feedback on the items, they answered a questionnaire evaluation form to assess their understanding and confidence in the standard setting process. Two specific questions were asked in this evaluation regarding the helpfulness of the item order and p-values:

- How helpful was the order of the items in the Booklet to your ratings?
- How helpful were the estimated p-values (item difficulty) to your Round 2 ratings?

The results of this evaluation were used in the overall comparison of the results.

5. Results

Table 2 gives a summary of the results of both groups and rounds using the raw score scale out of 50 items. The difference in the average recommended cut scores based on the means between the two groups represents less than 1 raw score point (0.6) for Round 1 and a difference of 2.21 raw score points for Round 2. The difference in the average recommended cut scores based on the medians between the two groups was 2 raw score points for Round 1 and a difference of a half raw score point for Round 2. The OIB Angoff group had a smaller standard deviation, range, and standard error of the mean than the Angoff group within both rounds.

As is expected with any multiple round standard setting, the variance among ratings within each group was reduced from Round 1 to Round 2. Figure 1 shows the recommended cut scores means as well as plus and minus two standard errors of the recommended mean cut score for each group for both rounds. This is one approach for modeling the variance observed in context to the average recommended cut scores. Although the size of the ranges varies between rounds and between methods, the bands are all overlapping with one another. This result suggests similar recommendations but with differing amounts of variability.
The consistency within each panelist’s ratings between the rounds was also of interest; these changes were evaluated in three ways. Table 3 shows (a) the average number of item-level ratings that were changed between rounds, (b) the average proportion of consistent item-level ratings between rounds, and (c) the average within person correlation of item-level ratings between rounds (Round 1 ratings correlated with Round 2 ratings).

The results in the table show that the members of the Angoff group, on average, made more changes to their item-level ratings between rounds (i.e., twice as many item-level ratings were changed). This is a logical finding considering the differences between the two methods. Round 2 was the first opportunity the Angoff panel had to use empirical estimates of difficulty in making their ratings whereas the OIB Angoff group had the relative difficulty information (booklet ordering) available during their first round of ratings.

In order to investigate the extent to which panelists disagreed with the item order as determined by the p-values (i.e., the order presented in the OIB), the rating of items before and after the cut score was evaluated for each panelist. Specifically, the recommended cut score by panel and round was used as a dividing line for all items when ordered based on difficulty. Any items before that dividing line that were rated as “No” and any items after that line rated as “Yes” were identified for each panelist. These counts indicate the level of disagreement that panelists had with the Guttman-like order of the items (i.e., all items rated as “No” up to a certain point and then all items rated as “Yes”). The median agreement and disagreement before and after the cut score for each panel and each round was computed and is shown in Figure 2. For example, in the first round of ratings, the Angoff panel identified 7 items (14% disagreement before the cut score) before the cut score that they felt an MCC would not likely answer correctly and 8 items (16% disagreement after the cut score) after the cut score that they felt the MCC would answer correctly. As shown in the table, the OIB Angoff group had slightly more of these items in both rounds of rating and both groups had more of these disagreements in Round 1 as compared to Round 2.

Finally, the panelists’ evaluation results were compared between panels regarding the helpfulness of the item order and the item-level p-values. The Angoff panel

---

### Table 2. Summary of Recommended cut score results by round

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angoff Group</td>
<td>OIB Angoff Group</td>
</tr>
<tr>
<td>Number of panelists</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Mean</td>
<td>36.90</td>
<td>37.50</td>
</tr>
<tr>
<td>Median</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>Minimum</td>
<td>24</td>
<td>31</td>
</tr>
<tr>
<td>Maximum</td>
<td>46</td>
<td>42</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>8.38</td>
<td>5.09</td>
</tr>
<tr>
<td>Standard Error of the Mean</td>
<td>3.17</td>
<td>2.08</td>
</tr>
</tbody>
</table>

### Table 3. Rating consistencies between rounds

<table>
<thead>
<tr>
<th></th>
<th>Angoff Group</th>
<th>OIB-Angoff Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of changed item ratings</td>
<td>10.14</td>
<td>4.67</td>
</tr>
<tr>
<td>Average proportion of consistent item-level ratings</td>
<td>0.80</td>
<td>0.91</td>
</tr>
<tr>
<td>Average within person correlation of item-level ratings</td>
<td>0.46</td>
<td>0.71</td>
</tr>
</tbody>
</table>

---

**Figure 1.** Recommended cut score ranges.

The consistency within each panelist’s ratings between the rounds was also of interest; these changes were evaluated in three ways. Table 3 shows (a) the average number of item-level ratings that were changed between rounds, (b) the average proportion of consistent item-level ratings between rounds, and (c) the average within person correlation of item-level ratings between rounds (Round 1 ratings correlated with Round 2 ratings).

The results in the table show that the members of the Angoff group, on average, made more changes to their item-level ratings between rounds (i.e., twice as many item-level ratings were changed). This is a logical finding considering the differences between the two methods. Round 2 was the first opportunity the Angoff panel had to use empirical estimates of difficulty in making their ratings whereas the OIB Angoff group had the relative difficulty information (booklet ordering) available during their first round of ratings.

In order to investigate the extent to which panelists disagreed with the item order as determined by the p-values (i.e., the order presented in the OIB), the rating of items before and after the cut score was evaluated for each panelist. Specifically, the recommended cut score by panel and round was used as a dividing line for all items when ordered based on difficulty. Any items before that dividing line that were rated as “No” and any items after that line rated as “Yes” were identified for each panelist. These counts indicate the level of disagreement that panelists had with the Guttman-like order of the items (i.e., all items rated as “No” up to a certain point and then all items rated as “Yes”). The median agreement and disagreement before and after the cut score for each panel and each round was computed and is shown in Figure 2. For example, in the first round of ratings, the Angoff panel identified 7 items (14% disagreement before the cut score) before the cut score that they felt an MCC would not likely answer correctly and 8 items (16% disagreement after the cut score) after the cut score that they felt the MCC would answer correctly. As shown in the table, the OIB Angoff group had slightly more of these items in both rounds of rating and both groups had more of these disagreements in Round 1 as compared to Round 2.

Finally, the panelists’ evaluation results were compared between panels regarding the helpfulness of the item order and the item-level p-values. The Angoff panel

---

### Table 2. Summary of Recommended cut score results by round

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angoff Group</td>
<td>OIB Angoff Group</td>
</tr>
<tr>
<td>Number of panelists</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Mean</td>
<td>36.90</td>
<td>37.50</td>
</tr>
<tr>
<td>Median</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>Minimum</td>
<td>24</td>
<td>31</td>
</tr>
<tr>
<td>Maximum</td>
<td>46</td>
<td>42</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>8.38</td>
<td>5.09</td>
</tr>
<tr>
<td>Standard Error of the Mean</td>
<td>3.17</td>
<td>2.08</td>
</tr>
</tbody>
</table>

### Table 3. Rating consistencies between rounds

<table>
<thead>
<tr>
<th></th>
<th>Angoff Group</th>
<th>OIB-Angoff Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of changed item ratings</td>
<td>10.14</td>
<td>4.67</td>
</tr>
<tr>
<td>Average proportion of consistent item-level ratings</td>
<td>0.80</td>
<td>0.91</td>
</tr>
<tr>
<td>Average within person correlation of item-level ratings</td>
<td>0.46</td>
<td>0.71</td>
</tr>
</tbody>
</table>

---

**Figure 1.** Recommended cut score ranges.
Combining the Best of Two Standard Setting Methods: The Ordered Item Booklet Angoff

Journal of Applied Testing Technology
Vol 15 (1) | 2014 | www.jattjournal.com

24

The results of the comparative standard setting panels in this study showed that the OIB Angoff group arrived at a recommended cut score that was not significantly different (less than 1 standard deviation) from the recommended cut score of the group using the traditional Angoff approach. The OIB Angoff group’s within group standard deviations, range, and standard errors show evidence of internal validity through the smaller within group variance and standard deviation during both rounds (Kane, 1994; 2001). Furthermore, this group’s ratings between rounds were far more consistent than the traditional Angoff group. In addition, there was some evidence to support the utility of the OIB when making the ratings (evaluation of presentation), which supports the idea that the OIB Angoff creates an easier task for panelists.

The patterns of ratings did not emerge as expected for the two standard setting panels. Following the logic of the Bookmark standard setting method, it was anticipated that the OIB Angoff group would essentially provide a Guttman-like pattern of ratings, with a clear distinction between items up to a certain point in the booklet that an MCC would answer correctly and then after which the MCC would answer incorrectly. This suggests that panelists do disagree with the order of items within an ordered item booklet; presenting the items from easiest to most difficult did not seem to influence the extent to which panelists agree with the expected Guttman ordering.

In considering the presentation of items, it should be noted that for the Angoff group, the items were clustered by content area within the booklet. Therefore, there could have been a mitigating factor that the panelists could judge all items from a given content area at once rather than having to switch topics from one item to the next as in the OIB Angoff group. To examine this, panelists were also asked to evaluate the helpfulness of the order of the items in making their ratings. The OIB Angoff group’s median and mode rating response was “very helpful” whereas the Angoff group’s median and mode rating response with unordered items was “helpful.” The positive perception from both group’s perspective suggests that there is value to ordering the items either by content area or by difficulty, and perhaps both.

While this study showed promise for the hybrid OIB Angoff standard setting method, the combination of the Angoff and Bookmark standard setting processes did not address every issue. With regard to the Bookmark method, one criticism not resolved by the OIB Angoff process is that the items must still be presented in a

Figure 2. Proportion of agreement and disagreement guttman scale relative to the cut score.

gave median and mode ratings for the item order as “helpful,” which was expectedly lower than the “very helpful” median and mode ratings for item order provided by the OIB Angoff group. The median and mode ratings for the helpfulness of the estimated p-values for both groups were “very helpful.”

6. Discussion

As discussed in previous research, the Angoff and Bookmark standard setting methods each have their strengths and weaknesses. The hybrid standard setting approach proposed in this article (OIB Angoff) attempts to address some—albeit not all—of the disadvantages of these individual standard setting methods. The OIB Angoff approach to standard setting is designed to reduce the cognitive challenge that panelists face when trying to make Angoff judgments (i.e., estimating item difficulty) by incorporating item difficulty information into the presentation of items while allowing panelists to disagree with the given order of items.

As a first investigation of this methodology, the OIB-Angoff was implemented concurrently with the Angoff method for a single cut score licensure examination. The intention of this study was to evaluate the feasibility of implementing this method and evaluate the outcomes in comparison to a well-tested method (Angoff). The example described in this article was executed using a Yes/No Angoff; however, the OIB Angoff could be implemented by asking panelists to estimate the likelihood that a target examinee would respond correctly.

The results of the comparative standard setting panels in this study showed that the OIB Angoff group arrived at a recommended cut score that was not significantly different (less than 1 standard deviation) from the recommended cut score of the group using the traditional Angoff approach. The OIB Angoff group’s within group standard deviations, range, and standard errors show evidence of internal validity through the smaller within group variance and standard deviation during both rounds (Kane, 1994; 2001). Furthermore, this group’s ratings between rounds were far more consistent than the traditional Angoff group. In addition, there was some evidence to support the utility of the OIB when making the ratings (evaluation of presentation), which supports the idea that the OIB Angoff creates an easier task for panelists.

The patterns of ratings did not emerge as expected for the two standard setting panels. Following the logic of the Bookmark standard setting method, it was anticipated that the OIB Angoff group would essentially provide a Guttman-like pattern of ratings, with a clear distinction between items up to a certain point in the booklet that an MCC would answer correctly and then after which the MCC would answer incorrectly. This suggests that panelists do disagree with the order of items within an ordered item booklet; presenting the items from easiest to most difficult did not seem to influence the extent to which panelists agree with the expected Guttman ordering.

In considering the presentation of items, it should be noted that for the Angoff group, the items were clustered by content area within the booklet. Therefore, there could have been a mitigating factor that the panelists could judge all items from a given content area at once rather than having to switch topics from one item to the next as in the OIB Angoff group. To examine this, panelists were also asked to evaluate the helpfulness of the order of the items in making their ratings. The OIB Angoff group’s median and mode rating response was “very helpful” whereas the Angoff group’s median and mode rating response with unordered items was “helpful.” The positive perception from both group’s perspective suggests that there is value to ordering the items either by content area or by difficulty, and perhaps both.

While this study showed promise for the hybrid OIB Angoff standard setting method, the combination of the Angoff and Bookmark standard setting processes did not address every issue. With regard to the Bookmark method, one criticism not resolved by the OIB Angoff process is that the items must still be presented in a

Figure 2. Proportion of agreement and disagreement guttman scale relative to the cut score.

gave median and mode ratings for the item order as “helpful,” which was expectedly lower than the “very helpful” median and mode ratings for item order provided by the OIB Angoff group. The median and mode ratings for the helpfulness of the estimated p-values for both groups were “very helpful.”

6. Discussion

As discussed in previous research, the Angoff and Bookmark standard setting methods each have their strengths and weaknesses. The hybrid standard setting approach proposed in this article (OIB Angoff) attempts to address some—albeit not all—of the disadvantages of these individual standard setting methods. The OIB Angoff approach to standard setting is designed to reduce the cognitive challenge that panelists face when trying to make Angoff judgments (i.e., estimating item difficulty) by incorporating item difficulty information into the presentation of items while allowing panelists to disagree with the given order of items.

As a first investigation of this methodology, the OIB-Angoff was implemented concurrently with the Angoff method for a single cut score licensure examination. The intention of this study was to evaluate the feasibility of implementing this method and evaluate the outcomes in comparison to a well-tested method (Angoff). The example described in this article was executed using a Yes/No Angoff; however, the OIB Angoff could be implemented by asking panelists to estimate the likelihood that a target examinee would respond correctly.

The results of the comparative standard setting panels in this study showed that the OIB Angoff group arrived at a recommended cut score that was not significantly different (less than 1 standard deviation) from the recommended cut score of the group using the traditional Angoff approach. The OIB Angoff group’s within group standard deviations, range, and standard errors show evidence of internal validity through the smaller within group variance and standard deviation during both rounds (Kane, 1994; 2001). Furthermore, this group’s ratings between rounds were far more consistent than the traditional Angoff group. In addition, there was some evidence to support the utility of the OIB when making the ratings (evaluation of presentation), which supports the idea that the OIB Angoff creates an easier task for panelists.

The patterns of ratings did not emerge as expected for the two standard setting panels. Following the logic of the Bookmark standard setting method, it was anticipated that the OIB Angoff group would essentially provide a Guttman-like pattern of ratings, with a clear distinction between items up to a certain point in the booklet that an MCC would answer correctly and then after which the MCC would answer incorrectly. This suggests that panelists do disagree with the order of items within an ordered item booklet; presenting the items from easiest to most difficult did not seem to influence the extent to which panelists agree with the expected Guttman ordering.

In considering the presentation of items, it should be noted that for the Angoff group, the items were clustered by content area within the booklet. Therefore, there could have been a mitigating factor that the panelists could judge all items from a given content area at once rather than having to switch topics from one item to the next as in the OIB Angoff group. To examine this, panelists were also asked to evaluate the helpfulness of the order of the items in making their ratings. The OIB Angoff group’s median and mode rating response was “very helpful” whereas the Angoff group’s median and mode rating response with unordered items was “helpful.” The positive perception from both group’s perspective suggests that there is value to ordering the items either by content area or by difficulty, and perhaps both.

While this study showed promise for the hybrid OIB Angoff standard setting method, the combination of the Angoff and Bookmark standard setting processes did not address every issue. With regard to the Bookmark method, one criticism not resolved by the OIB Angoff process is that the items must still be presented in a
particular order (by difficulty) and cannot be presented in any other order (e.g., by content area, exam order). However, further study of the OIB Angoff method could involve creating OIBs within each content area, which would be conceptually similar to the Mapmark method (Schulz & Mitzel, 2005). This design would minimize the potential analysis issues of Bookmark methods with separate content booklets (e.g., estimating a compensatory cut score using a conjunctive standard setting approach) because individual item ratings would still be collected. With regard to the Angoff method, the OIB Angoff process does not reduce the number of judgments for panelists.

A few key limitations in this study should be noted because they may restrict the generalizability of these results and prompt the need for future research. First, the sample size of the study was small—both with regard to the number of standard setting panelists (6 and 7, respectively) and the exam itself (a single form of a licensure exam). It is important to note that the results from this study may vary with more expanded standard setting panels and in different kinds of testing programs, particularly ones with a greater amount of variation in item difficulties. Second, the standard setting methodologies utilized in this study varied beyond the planned differences in the OIB. Specifically, the evaluation results are confounded by Angoff panelists’ additional discussions between rounds. This additional discussion may have resulted in a change in the confidence of the results. Third, one of the goals of implementing this methodology was to reduce the cognitive challenge faced by panelists when asked to rate the difficulty of items in the traditional Angoff implementations. However, the Angoff panel in this study was presented with the items grouped by content area, thus allowing panelists to focus on one area of knowledge and skills at a time. Although the evaluation results suggest that this ordering was not perceived to be as helpful as the OIB, additional research with items presented in random (e.g., test) order is needed.

7. Conclusions

The debate regarding the appropriateness and relative merit of various standard setting methods will continue as long as inferences regarding individuals’ knowledge, skills, and abilities are made using test scores. Given the importance of the resulting decisions from test scores, new standard setting methods seeking to enhance the strengths and diminish the weaknesses of existing processes will continue to populate the field. These methods must then be vetted by the measurement community against expectations for standard setting procedures and tested through implementation with various types of standard setting activities.

This study presents one such innovative approach by discussing the theoretical framework for a hybrid standard setting method combining the Bookmark and Angoff methods. It shows how the relevant discussion and empirical research on these popular standard setting methods has led to this logical combination of the strengths of these respective standard setting methods that simultaneously reduces some of their inherent weaknesses. When implemented consistently with published guidelines, this hybrid OIB Angoff approach should meet expectations for procedural validity within the measurement community because the use of an ordered item booklet within an Angoff standard setting does not greatly alter the generally accepted Angoff methods. Further, the hybrid methodology may actually make the standard setting task less cognitively complex for panelists.

As an initial look, the results from this study support the claim that the results from the OIB Angoff method appear to align with results from another commonly used method (e.g., similar recommended cut scores to the Angoff method) with some potential advantage to the panelists (e.g., less challenging task, smaller amounts of variance among panelists).

This study is only the first step in investigating the impact of using ordered item booklets in Angoff standard settings, but it does provide an initial indication of the potential utility of merging these methodologies. More research is needed to support or negate the assertion that this combined approach could reduce the cognitive burden of panelists’ rating items. This should continue to be investigated in different standard setting situations (within more diverse standard setting panels, different types of exams, and expanded ranges of item difficulty) and possibly through quantitative (e.g., study data) and qualitative (e.g., evaluation focus group, discussion with panelists) means.

Ultimately, there is not one best standard setting method that will be appropriate for all exams given policy, program design, and procedural considerations. However, through research efforts such as this study, attempts can be made to refine and improve existing methods by suggesting new modifications to make standard setting methods clearer and more coherent to panelists and thus,
produce more defensible cut score recommendations that align with panelists’ expectations.

8. Acknowledgement

The authors would like to acknowledge the Southern Research Testing Agency (SRTA) for their support in this research.

9. References


