Enhancements to the Bookmark and Item Descriptor Matching Standard Setting Methods

By Deborah L. Schnipke and Russell Keglovits

ACS Ventures, LLC

Abstract: In a typical Bookmark or Item Descriptor (ID) Matching standard setting meeting, subject matter experts (SMEs) rate the items on the criteria of the method, then use those ratings to determine performance standards (cutscores) that categorize test takers' scores into one of two or more categories (e.g., pass/fail, basic/proficient/advanced). During the course of a standard setting study, the authors of this paper found themselves in a situation where SMEs had no trouble with their item-level ratings but found it very challenging to determine performance standards based on their ratings. To help SMEs in such situations, the authors collected the item-level ratings (which is commonly not done when using these standard setting methods) and created visual displays to provide feedback to individual SMEs and also overall results to use for group feedback discussions. The authors also calculated empirical cutscores based on the item-level ratings for SMEs to use as a starting place for considering where to place their cutscores. By recording item-level ratings and providing feedback described in this paper, the potentially frustrating task of determining SME-level cutscores can be eased or eliminated, and instead SMEs can focus their time on rating the items, discussing group-level feedback, refining their item-level ratings, and coming to consensus on the group-level results and recommendations.

Introduction

Standard setting is a process of determining performance standards (cutscores) that categorize test takers' scores into one of two or more categories (e.g., pass/fail, basic/proficient/advanced). The performance standards are determined by collecting evidence-based judgments from a panel of well-qualified subject matter experts (SMEs) using one or more of a number of standard setting methods (Cizek & Bunch, 2007; Zieky et al., 2008).

In the bookmark (Karantonis & Sireci, 2006; Mitzel et al., 2001) and the item descriptor (ID) matching (Ferrara et al., 2014) standard setting methods, panelists review items sorted by difficulty, make judgments about the individual items, and provide cutscore(s) based on their string of judgments to delineate levels of performance. Both the bookmark and ID matching methods assume that a relatively clear pattern will result from the ratings and that panelists will be able to determine cutscores within their pattern of ratings. However, panelists sometimes have great difficulty with determining cutscores within their pattern of ratings, perhaps due to problem sensitivity, reasoning, and information ordering (Cizek et al., 2001). Zieky, Perie, and Livingston (2008) note that for the ID Matching method (and the criticism applies equally well to the Bookmark method), if no clear pattern emerges for most of the panelists, the method is not working, and a different cutscore method is required. Unfortunately, changing methods in the middle of a standard setting meeting is not feasible.

However, it is feasible to collect item-level ratings (which is not typically done) that can be used to 1) provide additional types of feedback to panelists to help them determine cutscores within their pattern of ratings and/or 2) eliminate the need to have panelists determine cutscores within their pattern of ratings by simply counting the number of ratings in each category and use that as the panelist-level cutscores.

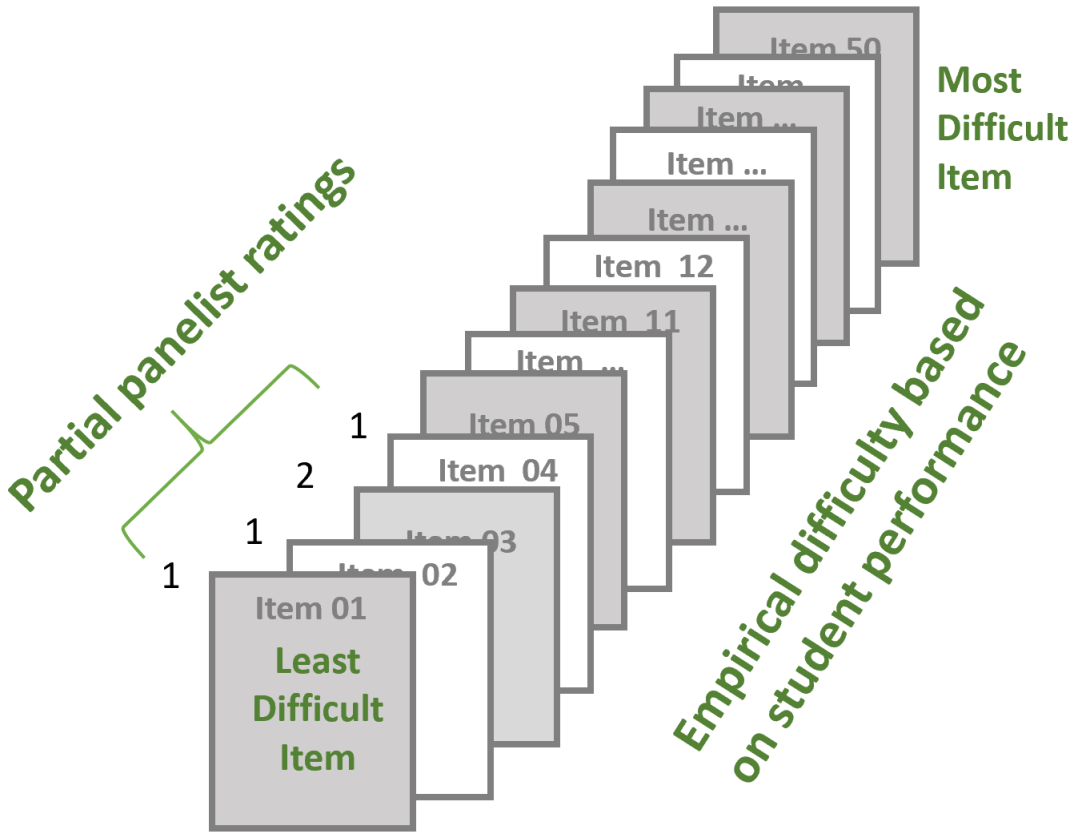
The authors of this paper found themselves in the situation of unclear patterns during a recent ID matching method standard setting meeting. Panelists were able to confidently make item-level judgments but found it extremely difficult, time consuming, and frustrating to have to choose cutscores based on their ratings. Due to the complaints of the panelists, the facilitators decided to gather the item-level ratings (which was not initially planned) and investigate ways of helping the panelists determine their cutscores.

In this paper, we first discuss a typical bookmark of ID matching standard setting process, then we discuss enhancements to the process using item-level data.

Typical Bookmark or ID Matching Process

In a typical bookmark or ID matching standard setting meeting, panelists are provided with an ordered item booklet which displays the items from easiest to hardest based on empirical results of student performance, as represented in Figure 1. The panelists read through the items to determine when the items switch from being mostly of the first category (based on expected student performance in the bookmark method or based on how well the item content matches the achievement level descriptors for the items in the ID matching method) to being mostly of the second category, and from the second to the third category, from third to fourth category, and so on for as many categories there are. They might decide on a threshold region within which each transition occurs, or a specific cutscore where they believe the transition occurs. When they are finished, the facilitator will have them submit either the beginning and end of each threshold region (from which the mean is usually calculated), or their individual cutscores to indicate these transitions.

Figure 1. Representation of ordered item booklet used to rate item in bookmark and ID matching standard setting



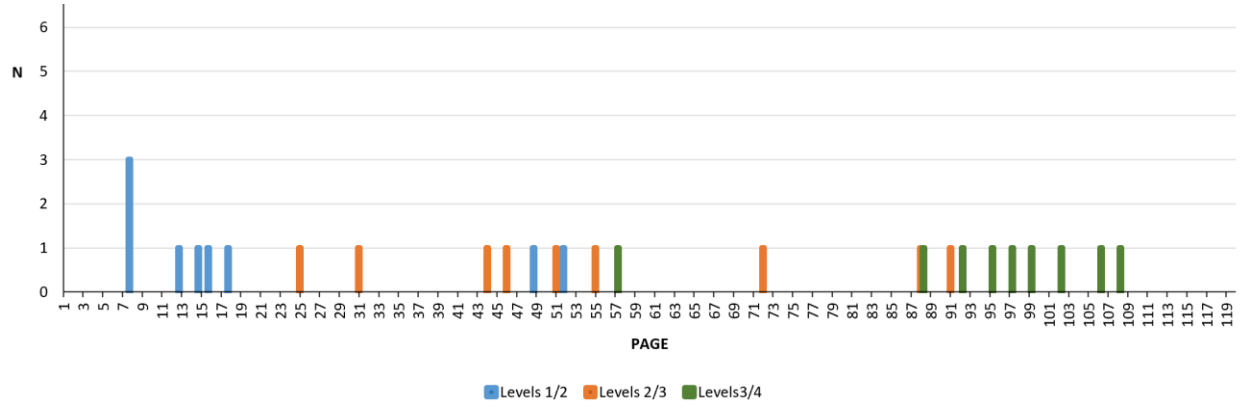
Both the Bookmark and ID Matching methods rely on the panelists being able to establish appropriate cutscores based on the ordered items. Panelists typically categorize items as they progress through the ordered items (writing their judgments on the ordered item booklet, a piece of scrap paper, or ideally on a rating sheet), then review their own ratings to determine where to place their cutscores. For example, assume there will be two cutscores, delineating groups 1, 2, and 3 (e.g., into basic, proficient, and advanced). In Bookmark standard setting, panelist rate the items based on whether the first borderline students (between categories 1 and 2) would "probably" (e.g., two-thirds of the time) be able to answer able to answer the question correctly (or get the score point) (coded 1 in this example). If the question is too difficult for the 1-2 borderline students, it is coded 2 signifying that the 2-3 borderline students are more likely to get the score point. If the question is too difficult for the 2-3 borderline students, it is coded 3. Because the items are ordered from easiest to hardest, it is expected that items will mostly be coded as 1s at first, followed by mostly 2s, and so on, as in this example: 11111211222122222223223332333333333. Threshold regions, underlined in the example, are defined by the mixed pattern of numbers. Each panelist will have his or her own string of ratings and must determine where to place their cutscores within their threshold regions. Typically, only the cutscores (rather than the string of ratings) are provided to the workshop facilitator. The facilitator compiles the results across panelists, calculates various statistics such as the median cutscore for each level, and provides group-level feedback to panelists which they discuss. Panelists then rerate the items (round 2), typically see more feedback, and do final (round 3) ratings. At the end, the median of the cutscores for each level from the final round of ratings is recommended as the overall cutscores.

ID matching is similar to the bookmark method, but it is distinguished by the framework against which student performance is conceptualized. In ID matching, the item is matched to achievement level descriptors (ALD) rather than borderline students. ID matching is presented as being less cognitively challenging for SME in so much as they are not required to think in terms of the likelihood of a correct response; rather, they are to match the item to a descriptor. These descriptors range from the lowest to highest level of cognitive rigor for the student. In the ID Matching method, ratings are based on whether the item best matches the lowest performance category (1), the next highest next highest category (2), and so on, resulting in a pattern that looks like a pattern resulting from the Bookmark method. Although the ratings arise from a different rating process, the process of the panelists of determining their threshold regions and placing their cutscores in their threshold regions is the same as in the Bookmark method.

Group-level is used in Bookmark or ID Matching standard setting meetings to help the panelists reconcile the differences between panelists and to make a final recommendation for each cutscore. Typical group-level feedback in a Bookmark or ID Matching standard setting meeting includes:

- Min, max, and median of panelists cutscores for each level
- Graph of individual cutscores, color-coded by level (such as in Figure 2)
- Impact data (the percentage of students who would be classified into each category based on the median cutscore for each level)

Figure 2. Sample group-level feedback in typical a Bookmark or ID Matching standard setting meeting



Using the group-level feedback, panelists compare their own cutscores to the group results, consider the impact of their cutscores on the test taker population, and reconsider their ratings and cutscores. The group discusses the overall results after several rounds of ratings and comes to a group consensus/recommendation.

Both the bookmark and ID matching standard setting methods work quite well when the panelists can detect a reasonable progression in their ratings, allowing them to place their cutscores within the threshold regions in a relatively straightforward manner. They might be encouraged to choose the middle of the threshold range, for example. However, if the series of ratings do not fall into a clear pattern, panelists struggle with choosing their cutscores. For example, this is a string of ratings from an ID matching standard setting with four categories (3 cutscores) performed by the authors: 111131231112231222121232213223221222222232432311412232222. This panelist was uncertain about where to place the transition regions, and was challenged to identify specific cutscores, but finally decided on the underlined numbers as the cutscores. The panelist was frustrated and not confident about the cutscores. Other panelists had similar difficulties. The enhanced process described below reviews changes we made (or wanted to make) to the process and will be standard practice for us in future standard setting meetings.

Enhanced Feedback Using Item-Level Data

Because panelists will be making item-level judgments to determine their cutscores in a bookmark or ID matching standard setting meeting, it is possible to capture those ratings and make use of them. Whether panelist enter their item-level ratings in a spreadsheet or on paper, if the facilitator and/or assistant enters all panelists item-level ratings into a master spreadsheet, additional group-level feedback can be provided to help the panelists evaluate the items and determine the cutscores.

This section uses data from an ID Matching standard setting recently performed by the authors. On paper rating sheets, panelists were asked to mark one of four performance levels to which each item mapped. They could indicate 1.5, 2.5, or 3.5 if desired when the performance level was judged to be between two categories. Panelists were able to do the rating task without much difficulty. They were then instructed to find the threshold region were the ratings switched from one achievement level to the next, and all panelists struggled with this task. The threshold region could be a single page, or a range of pages in the ordered item booklet. Although the original plan was to only enter their threshold regions in the master spreadsheet, the facilitators decided to enter the item-level ratings as well to see if they could make better sense of the ratings and provide additional help to the panelists.

Due to the last-minute nature of the changes, Figure 3 was the one used during the meeting because it can be created from scratch relatively quickly. All item-level ratings were entered into a master spreadsheet as they were handed in (and during a break before feedback discussions), and conditional formatting was applied. Although many of the columns of data look messy individually, a pattern starts to emerge when viewing them as a group. The last column shows the median rating, and highlighting in that column shows the mean and median ratings for each cutscore (orange for the cutscore between level 1 and 2, yellow for level 2 to 3, and green for level 3 to 4). There were three rounds of ratings, and in the third round, no changes were made to the item-level ratings, but the panelists did change their thresholds, and that is shown in a column marked Round 3 in the graph in the right panel.

What the authors think would have been better is shown in Figure 4, which shows the same ratings, but with a different visual display. This display requires either setting up the file in advance to apply conditional formatting based on rules that incorporate the minimum and maximum thresholds provided by the panelists (recommended) or applying the color coding manually during the meeting (not recommended due to this approach being slower and more error prone). The threshold regions provided by the panelists are in Figure 4 for all three rounds of the process (round 1 on the left, round 2 in the middle, and round 3 on the right). Orange indicates the threshold region between performance levels 1 and 2, blue indicates 2 to 3, and green indicates 3 to 4. Privately telling each panelist which column is their data (and not telling anyone which columns belong to the other participants) allows panelists to see very clearly how their threshold region compares to the group in both width and location and allows them to use richer data than Figure 2 to re-evaluate their item-level ratings and threshold regions.

Figure 3. Round 1 – 2 Ratings with conditional formatting and Mean and Median Cutscores in final column

Round 1 Ratings and Cutscores

Page	SME1	SME2	SME3	SME4	SME5	SME6	SME7	SME8	SME9	Median
1	1	2	2	2	2	2	2	1	2	2
2	1	2	2	2	2	1	1.5	1	2	2
3	2	2	2	2	2	1	2	1	2	2
4	1	1.5	1.5	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1.5	1	1	1	1
7	1	1	1	1	1	1.5	1	1	1	1
8	3	3	3	3	4	3	1	1.5	3.5	3
9	1	3	3	3	3	2	3	2	3	3
10	2	2	1	1	1	2	1	1	1	1
11	3	3	2	3	2	2.5	1.5	2	2	2
12	1	2	2	2	2	2.5	2	2	2	2
13	1	2	3	3	3	2.5	3	2	3	3
14	1	2	2	2	2	2.5	2	2	2	2
15	2	3	3	3	3	3	3	2.5	3	3
16	2	1	1.5	2	1	1	1	2.5	2	1.5
17	3	3	3	3	3	3	1.5	3	3	3
18	1	2	1	2.5	2	3	1	2	2	2
19	2	2.5	3.5	4	4	3.5	3	3	3	3
20	2	2	3	3	3	3	3	3	3	3
21	1	1	1	1	1	1	1	1	1	1
22	2	3	3.5	3.5	4	4	3	3	3.5	3.5
23	1	1	1	1.5	1	1	1	1	1	1
24	2	3	3	3.5	3	3.5	3.5	3	3.5	3
25	3	2	2.5	4	4	3.5	3	3	3	3
26	2	2	1.5	1.5	2	2	1.5	2	3	2
27	2	2	2	3	3	3.5	2.5	2	3	2.5
28	1	1	1	1	1	1	1	1	1	1
29	3	2	2	3	2	3	2	3	2	2
30	3	4	4	4	4	4	4	3	4	4
31	2	2	2	3	3	3	2	3	3	3
32	2	3	4	4	3	4	4	3	4	4
33	3	4	3	4	4	4	4	3	3.5	4
34	2	2	2	2	2	3	2	2	1	2
35	1	1	1	1	1	1	1	1	1	1
36	2	2.5	3	3	1	3	2.5	3	3	3
37	2	3	3	3	3	3.5	1	3	1	3
38	2	1	1	1	1	1	1	1	1	1
39	2	2	2	2	2	2	2	2	2	2
40	2	2	1	2	2	2	2	2	2	2
41	2	3	3	3	3	3	3	3	3	3
42	2	3	3	4	4	4	3	3.5	4	3.5
43	3	3	3	3	3	3.5	3	3	3	3
44	2	3	3	3.5	3.5	3.5	4	3.5	4	3.5
45	4	3	3	3	3	3.5	3	3	3	3
46	3	3	1.5	1	1.5	2.5	1	2.5	4	2.5
47	2	2	2	3	3	4	3	3	2	3
48	3	4	4	4	4	4	4	4	4	4
49	1	2	2	2	2	2	2	3	2	2
50	1	2	2	2	2	2	2	2	2	2
51	4	4	3	3.5	3	3.5	3	4	4	3.5
52	1	2	1	1.5	2	1	2	3	2	2
53	2	3	3	3	3	3	3	3.5	3	3
54	2	3.5	2	3	1	3	3	3	3	3
55	3	4	4	4	3	4	4	4	4	4
56	2	3	3	3	3.5	4	3	3	2	3
57	2	3	3.5	3.5	3.5	4	3	4	3	3.5
58	2	3	3	4	3.5	4	3	4	4	3.5

Round 2 Ratings and Cutscores (with Round 3 Cutscores)

Page	SME1	SME2	SME3	SME4	SME5	SME6	SME7	SME8	SME9	Median	Round 3
1	1	2	2	2	2	2	1	1	1	2	
2	1	1	2	1	1	2	1	1	1	1	
3	2	2	2	1	1	2	1	1	1	1	
4	1	1	1	1	1	1	1	1	1.5	1	
5	1	1	1	1	1	1	1	1	1	1	
6	1	1	1	1	1	1	1	1	1	1	
7	1	1	1	1	1	1	1	1	1	1	
8	2	2	2	1	3	4	4	1	1	2	
9	1.5	2	2	1.5	2	3	1.5	2.5	1	2	
10	1.5	1	1	1	1	1	1.5	1.5	1.5	1	Round 3
11	1.5	2	2	1.5	2	2	2	1.5	1.5	2	
12	2	2	2	1.5	2	2	2	2	1	2	
13	2	2	3	2	2	3	2	2	2	2	
14	2	1	2	1.5	2	2	2	2	1.5	2	
15	3	2	3	3	3	1	2	2.5	2	2.5	
16	2	1	2	2	2	1	2	2.5	1.5	2	
17	2	2	3	2	2	3	2	2.5	2.5	2	
18	2	3	2	2	2	2	2	2.5	2	2	
19	2	2	3	3	3.5	4	1	3	1	3	
20	3	1	3	2.5	3	3	2	3	3	3	
21	1	2	1	1	1	1	1	1	1	1	
22	2	2	3	2.5	3	3.5	3	2	3	3	
23	2	1	1.5	1	1	1	1	1	1	1	
24	3	2	3.5	2.5	3	3.5	3	3	3	3	
25	2	2	3	2.5	2	2.5	3	3	4	2.5	
26	2.5	2	1.5	2.5	3	1	2	3	3	2.5	
27	2	2	2.5	2.5	3	3	2	3	2	2.5	
28	1	1	1	2.5	1	1	1	1	1	1	
29	2	2	2.5	2.5	3	2	2.5	3	2	2.5	
30	3	3	3	3	4	4	3	3	4	3	Round 3
31	3	2	3	3	3	3	3	3	1	3	
32	3	2	3	4	4	4	3	3	2	3	
33	3	3	3.5	3	4	3	1	3	3.5	3	
34	3	2	2	3	2.5	2	2	2	1	2	
35	2	1	2	1	1	1	1	3	1	1	
36	3	2	2	3	3	2.5	3	3	3	3	Round 3
37	4	2	3	3	3.5	4	2	2	4	3	
38	2	1	1.5	2.5	1	1	1	2	1	1	
39	3	2	2	3	3	2	2	3	2	2	
40	2	2	2	2	2	2	2	1	1	2	
41	3	3	3	3	3	3	3	3	3	3	
42	3	3	3.5	3	4	4	4	3.5	3	3.5	
43	3	3	3	3	3	3	3	3.5	3	3	
44	4	3	3	3.5	4	4	3	3.5	3	3.5	
45	3	3	3	3	3	3	3	3.5	3	3	Round 3
46	3	2	2.5	3	3	1	2	3.5	3	3	
47	3	3	2.5	3.5	3	3	2	3.5	1.5	3	
48	4	3	3.5	3.5	4	4	3	3.5	4	3.5	
49	3	2	2	3.5	2	2	2	2	2	2	Round 3
50	3	2	2.5	3	2.5	2	2	3.5	2	2.5	
51	4	3	3.5	3.5	4	4	4	3.5	4	4	
52	3	2	2	3	3	1	2	3.5	2	2	
53	3	3	3	3	3	3	3	4	2	3	
54	3.5	3	2.5	3	3	3	3	3	3	3	
55	3.5	4	4	4	4	4	3.5	4	4	4	
56	3	4	3	4	3	3	3.5	4	2	3	
57	4	3	3.5	4	3.5	3	4	3	3	3.5	
58	4	4	3.5	4	4	2	4	4	4	4	

Figure 4. Round 1 – 3 Ratings with SME Threshold Regions (shaded)

Round 1

Page	SME1	SME2	SME3	SME4	SME5	SME6	SME7	SME8	SME9	Median
1	1	2	2	2	2	2	2	1	2	2
2	1	2	2	2	2	1	1.5	1	2	2
3	2	2	2	2	2	1	2	1	2	2
4	1	1.5	1.5	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1.5	1	1	1	1
7	1	1	1	1	1	1.5	1	1	1	1
8	3	3	3	3	4	3	1	1.5	3.5	3
9	1	3	3	3	3	2	3	2	3	3
10	2	2	1	1	1	2	1	1	1	1
11	3	3	2	3	2	2.5	1.5	2	2	2
12	1	2	2	2	2	2.5	2	2	2	2
13	1	2	3	3	3	2.5	3	2	3	3
14	1	2	2	2	2	2.5	2	2	2	2
15	2	3	3	3	3	3	3	2.5	3	3
16	2	1	1.5	2	1	1	1	2.5	2	1.5
17	3	3	3	3	3	3	1.5	3	3	3
18	1	2	1	2.5	2	3	1	2	2	2
19	2	2.5	3.5	4	4	3.5	3	3	3	3
20	2	2	3	3	3	3	3	3	3	3
21	1	1	1	1	1	1	1	1	1	1
22	2	3	3.5	3.5	4	4	3	3	3.5	3.5
23	1	1	1	1.5	1	1	1	1	1	1
24	2	3	3	3.5	3	3.5	3.5	3	3.5	3
25	3	2	2.5	4	4	3.5	3	3	3	3
26	2	2	1.5	1.5	2	2	1.5	2	3	2
27	2	2	2	3	3	3.5	2.5	2	3	2.5
28	1	1	1	1	1	1	1	1	1	1
29	3	2	2	3	2	3	2	3	2	2
30	3	4	4	4	4	4	4	3	4	4
31	2	2	2	3	3	3	2	3	3	3
32	2	3	4	4	3	4	4	3	4	4
33	3	4	3	4	4	4	4	3	3.5	4
34	2	2	2	2	2	3	2	2	1	2
35	1	1	1	1	1	1	1	1	1	1
36	2	2.5	3	3	1	3	2.5	3	3	3
37	2	3	3	3	3	3.5	1	3	1	3
38	2	1	1	1	1	1	1	1	1	1
39	2	2	2	2	2	2	2	2	2	2
40	2	2	1	2	2	2	2	2	2	2
41	2	3	3	3	3	3	3	3	3	3
42	2	3	3	4	4	4	3	3.5	4	3.5
43	3	3	3	3	3	3.5	3	3	3	3
44	2	3	3	3.5	3.5	3.5	4	3.5	4	3.5
45	4	3	3	3	3	3.5	3	3	3	3
46	3	3	1.5	1	1.5	2.5	1	2.5	4	2.5
47	2	2	2	3	3	4	3	3	2	3
48	3	4	4	4	4	4	4	4	4	4
49	1	2	2	2	2	2	2	3	2	2
50	1	2	2	2	2	2	2	2	2	2
51	4	4	3	3.5	3	3.5	3	4	4	3.5
52	1	2	1	1.5	2	1	2	3	2	2
53	2	3	3	3	3	3	3	3.5	3	3
54	2	3.5	2	3	1	3	3	3	3	3
55	3	4	4	4	3	4	4	4	4	4
56	2	3	3	3	3.5	4	3	3	2	3
57	2	3	3.5	3.5	3.5	4	3	4	3	3.5
58	2	3	3	4	3.5	4	3	4	4	3.5

Round 2

Page	SME1	SME2	SME3	SME4	SME5	SME6	SME7	SME8	SME9	Median
1	1	2	2	2	2	2	1	1	1	2
2	1	2	2	1	1	2	1	1	1	1
3	2	2	2	1	1	2	1	1	1	1
4	1	1.5	1	1	1	1	1	1	1.5	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1
8	2	3	2	1	3	4	4	1	1	2
9	1.5	3	2	1.5	2	3	1.5	2.5	1	2
10	1.5	2	1	1	1	1	1.5	1.5	1.5	1.5
11	1.5	3	2	1.5	2	2	2	1.5	1.5	2
12	2	2	2	1.5	2	2	2	2	1	2
13	2	2	3	2	2	3	2	2	2	2
14	2	2	2	1.5	2	2	2	2	1.5	2
15	3	3	3	3	3	1	2	2.5	2	3
16	2	1	2	2	2	1	2	2.5	1.5	2
17	2	3	3	2	2	3	2	2.5	2.5	2.5
18	2	2	2	2	2	2	2	2.5	2	2
19	2	2.5	3	3	3.5	4	1	3	1	3
20	3	2	3	2.5	3	3	2	3	3	3
21	1	1	1	1	1	1	1	1	1	1
22	2	3	3	2.5	3	3.5	3	2	3	3
23	2	1	1.5	1	1	1	1	1	1	1
24	3	3	3.5	2.5	3	3.5	3	3	3	3
25	2	2	3	2.5	2	2.5	3	3	4	2.5
26	2.5	2	1.5	2.5	3	1	2	3	3	2.5
27	2	2	2.5	2.5	3	3	2	3	2	2.5
28	1	1	1	2.5	1	1	1	1	1	1
29	2	2	2.5	2.5	3	2	2.5	3	2	2.5
30	3	4	3	3	4	4	3	3	4	3
31	3	2	3	3	3	3	3	3	3	3
32	3	3	3	4	4	4	3	3	2	3
33	3	4	3.5	3	4	3	1	3	3.5	3
34	3	2	2	3	2.5	2	2	2	1	2
35	2	1	2	2	1	1	1	1	1	1
36	3	2.5	2	3	3	2.5	3	3	3	3
37	4	3	3	3	3.5	4	2	2	4	3
38	2	1	1.5	2.5	1	1	1	2	1	1
39	3	2	2	3	3	2	2	3	2	2
40	2	2	2	2	2	2	2	1	1	2
41	3	3	3	3	3	3	3	3	3	3
42	3	3	3.5	3	4	4	4	3.5	3	3.5
43	3	3	3	3	3	3	3	3.5	3	3
44	4	3	3	3.5	4	4	3	3.5	3	3.5
45	3	3	3	3	3	3	3	3.5	3	3
46	3	3	2.5	3	3	1	2	3.5	3	3
47	3	2	2.5	3.5	3	3	2	3.5	1.5	3
48	4	4	3.5	3.5	4	4	3	3.5	4	4
49	3	2	2	3.5	2	2	2	2	2	2
50	3	2	2.5	3	2.5	2	2	3.5	2	2.5
51	4	4	3.5	3.5	4	4	4	3.5	4	4
52	3	2	2	3	3	1	2	3.5	2	2
53	3	3	3	3	3	3	3	4	2	3
54	3.5	3.5	2.5	3	3	3	3	3	3	3
55	3.5	4	4	4	4	4	3.5	4	4	4
56	3	4	3	4	3.5	4	3	4	2	3
57	4	3	3.5	4	3.5	3	4	3	3	3.5
58	4	3	3.5	4	4	2	4	4	4	4

Round 3

Page	SME1	SME2	SME3	SME4	SME5	SME6	SME7	SME8	SME9	Median
1	1	2	2	2	2	2	1	1	1	2
2	1	1	2	1	1	2	1	1	1	1
3	2	2	2	1	1	2	1	1	1	1
4	1	1	1	1	1	1	1	1	1.5	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1
8	2	2	2	1	3	4	4	1	1	2
9	1.5	2	2	1.5	2	3	1.5	2.5	1	2
10	1.5	1	1	1	1	1	1.5	1.5	1.5	1
11	1.5	2	2	1.5	2	2	2	1.5	1.5	2
12	2	2	2	1.5	2	2	2	2	1	2
13	2	2	3	2	2	3	2	2	2	2
14	2	1	2	1.5	2	2	2	2	1.5	2
15	3	2	3	3	3	1	2	2.5	2	2.5
16	2	1	2	2	2	1	2	2.5	1.5	2
17	2	2	3	2	2	3	2	2.5	2.5	2
18	2	3	2	2	2	2	2	2.5	2	2
19	2	2	3	3	3.5	4	1	3	1	3
20	3	1	3	2.5	3	3	2	3	3	3
21	1	2	1	1	1	1	1	1	1	1
22	2	2	3	2.5	3	3.5	3	2	3	3
23	2	1	1.5	1	1	1	1	1	1	1
24	3	2	3.5	2.5	3	3.5	3	3	3	3
25	2	2	3	2.5	2	2.5	3	3	4	2.5
26	2.5	2	1.5	2.5	3	1	2	3	3	2.5
27	2	2	2.5	2.5	3	3	2	3	2	2.5
28	1	1	1	2.5	1	1	1	1	1	1
29	2	2	2.5	2.5	3	2	2.5	3	2	2.5
30	3	3	3	3	4	4	3	3	4	3
31	3	2	3	3	3	3	3	3	1	3
32	3	2	3	4	4	4	3	3	2	3
33	3	3	3.5	3	4	3	1	3	3.5	3
34	3	2	2	3	2.5	2	2	2	1	2
35	2	1	2	2	1	1	1	1	1	1
36	3	2	2	3	3	2.5	3	3	3	3
37	4	2	3	3	3.5	4	2	2	4	3
38	2	1	1.5	2.5	1	1	1	2	1	1
39	3	2	2	3	3	2	2	3	2	2
40	2	2	2	2	2	2	2	1	1	2
41	3	3	3	3	3	3	3	3	3	3
42	3	3	3.5	3	4	4	4	3.5	3	3.5
43	3	3	3	3	3	3	3	3.5	3	3
44	4	3	3	3.5	4	4	3	3.5	3	3.5
45	3	3	3	3	3	3	3	3.5	3	3
46	3	2	2.5	3	3	1	2	3.5	3	3
47	3	3	2.5	3.5	3	3	2	3.5	1.5	3
48	4	3	3.5	3.5	4	4	3	3.5	4	3.5
49	3	2	2	3.5	2	2	2	2	2	2
50	3	2	2.5	3	2.5	2	2	3.5	2	2.5
51	4	3	3.5	3.5	4	4	4	3.5	4	4
52	3	2	2	3	3	1	2	3.5	2	2
53	3	3	3	3	3	3	3	4	2	3
54	3.5	3	2.5	3	3	3	3	3	3	3
55	3.5	4	4	4	4	4	3.5	4	4	4
56	3	4	3	4	3	3	3.5	4	2	3
57	4	3	3.5	4	3.5	3	4	3	3	3.5
58	4	4	3.5	4	4	2	4	4	4	4

Immediate Panelist Feedback

If it is feasible to have panelists enter their item-level data into a spreadsheet during the rating task (as opposed to writing them on paper), immediate individual-level feedback can be provided to panelists to help them choose their individual cutscores (or to eliminate the need to choose them at all).

Figure 5 shows an example of immediate item-level feedback for two different panelists. On the left is a panelist with "clean" data (from a different exam than Figures 3-4), and on the right is a panelist with "messy" data (SME1 from Figures 3-4). For a panelist with "clean" ratings that match the expected order of ratings reasonably well (i.e., mostly 1s, followed by mostly 2s, followed by mostly 3s, etc.), the feedback provided by the color coding and count of ratings by category may be helpful, although such panelists probably wouldn't have too much trouble selecting their individual cutscores. For panelists with "messy" data that do not show a clear pattern, the immediate feedback provided by the spreadsheet can provide valuable information to help them decide where to place their threshold regions. Although it may be difficult for the SME with messy data in Figure 5 (right side) to choose cutscores, it is hoped that the feedback will make it easier for them. The panelists may use the potential cutscores that are based solely on the number of ratings of each type, or they may decide those potential cutscores are inappropriate and choose to move them up or down. The facilitator might suggest to the panelists that if they cannot decide, then they could use the potential cutscores for that round of ratings.

To create Figure 5, set up conditional formatting in advance so that as panelists enter their ratings, the color coding is applied. In addition, set up a table that counts the number of each rating and the cumulative count. The number of ratings for category 1 is the potential cutscore between categories 1 and 2. The number of ratings for categories 1 and 2 combined is the potential cutscore between categories 2 and 3. The number of ratings for categories 1, 2, and 3 combined is the potential cutscore between categories 3 and 4. Set up the spreadsheet so that when the cumulative total matches the total test length (58 in Figure 3), a label such as "Potential cutscore 1" appears next to that item in the item rating list.

Figure 5. Sample ratings for two panelists including color-coding based on the rating and potential cutscores based on cumulative counts

"Clean" data (different exam than Figures 3-4)

OIB #	SME12		Rating	Number	Cum.Num	
1	1		1	8	8	Potential cutscore 1
2	1		2	27	35	Potential cutscore 2
3	1		3	11	46	Potential cutscore 3
4	1					
5	1					
6	1					
7	1					
8	2	Potential cutscore 1				
9	1					
10	2					
11	2					
12	2					
13	2					
14	2					
15	2					
16	2					
17	2					
18	2					
19	2					
20	2					
21	2					
22	2					
23	2					
24	2					
25	2					
26	2					
27	2					
28	2					
29	2					
30	2					
31	2					
32	2					
33	2					
34	3					
35	2	Potential cutscore 2				
36	3					
37	3					
38	3					
39	3					
40	3					
41	3					
42	3					
43	3					
44	4					
45	3					
46	4	Potential cutscore 3				
47	2					
48	3					
49	4					
50	4					
51	4					
52	4					
53	4					
54	4					
55	4					
56	4					
57	4					
58	4					

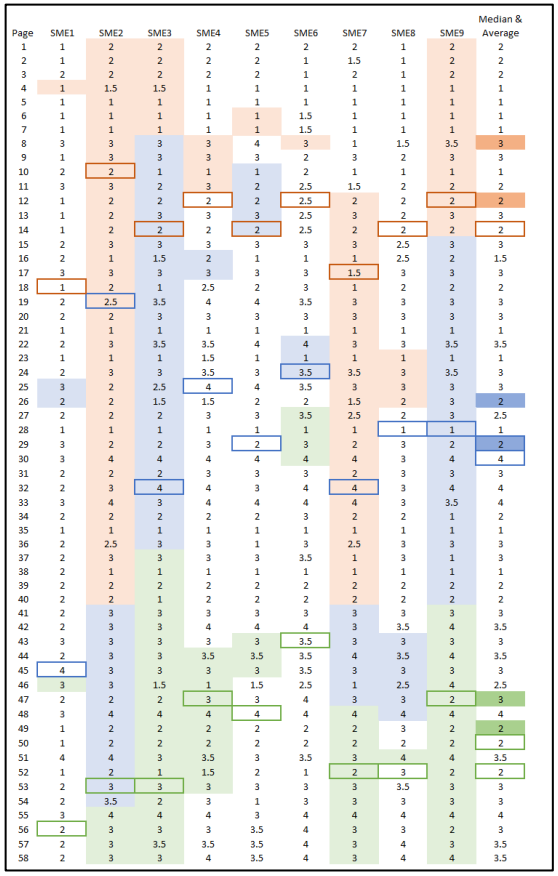
"Messy" data (SME01 from Figures 3-4)

OIB #	SME01		Rating	Number	Cum.Num	
1	1		1	18	18	Potential cutscore 1
2	1		2	27	45	Potential cutscore 2
3	2		3	11	56	Potential cutscore 3
4	1					
5	1					
6	1					
7	1					
8	3					
9	1					
10	2					
11	3					
12	1					
13	1					
14	1					
15	2					
16	2					
17	3					
18	1	Potential cutscore 1				
19	2					
20	2					
21	1					
22	2					
23	1					
24	2					
25	3					
26	2					
27	2					
28	1					
29	3					
30	3					
31	2					
32	2					
33	3					
34	2					
35	1	Potential cutscore 2				
36	2					
37	2					
38	2					
39	2					
40	2					
41	2					
42	2					
43	3					
44	2					
45	4					
46	3					
47	2					
48	3					
49	1					
50	1					
51	4					
52	1					
53	2					
54	2					
55	3					
56	2	Potential cutscore 3				
57	2					
58	2					

The potential cutscores in Figure 5 can also be applied to the data in Figure 4, resulting in Figure 6. The potential cutscores for each panelist are shown with a box around the rating in the row for each potential cutscore, color coded to match the thresholds provided by the panelists. Figure 6 was not shown to the panelists in the study, but we believe the feedback discussions would have been richer, smoother, and less frustrating with feedback such as this, and the results may have looked cleaner as well. We plan to implement individual feedback as in Figure 5 and group feedback as in Figure 5 followed by Figure 6 after each round (shown only graph for that round rather than all 3 rounds at once) at our next opportunity.

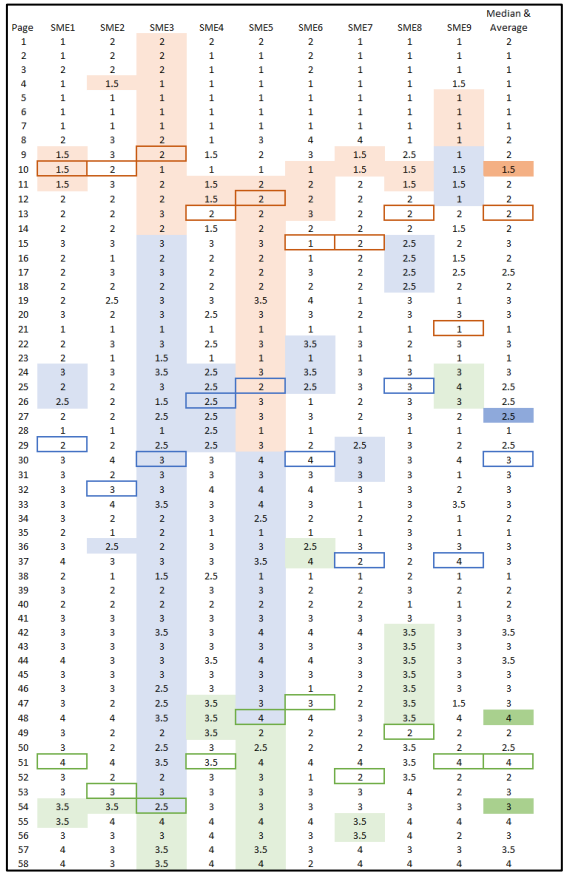
Figure 6. Round 1 – 3 Ratings with SME Threshold Regions (shaded) and Estimated Cutscores Based on Item-Level Ratings (borders)

Round 1



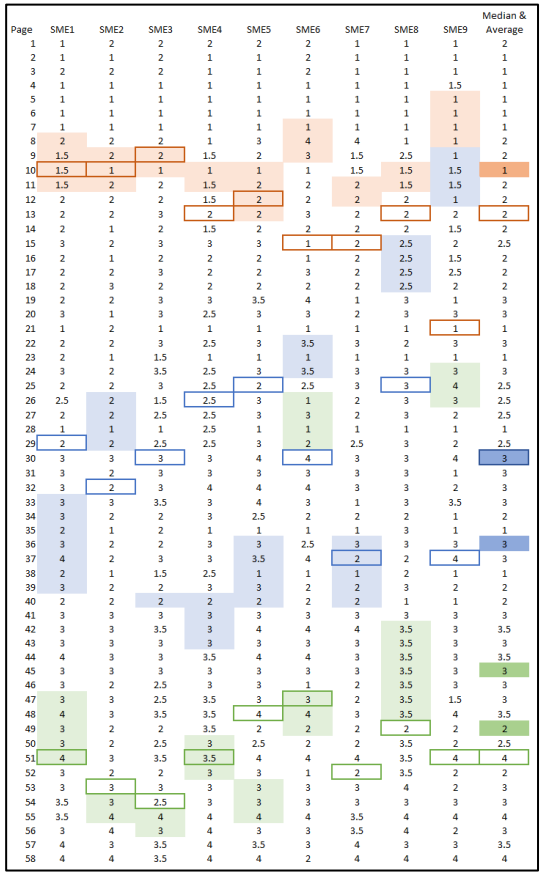
Page	SME1	SME2	SME3	SME4	SME5	SME6	SME7	SME8	SME9	Median & Average
1	1	2	2	2	2	2	2	1	2	2
2	1	2	2	2	2	1	1.5	1	2	2
3	2	2	2	2	2	1	2	1	2	2
4	1	1.5	1.5	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1.5	1	1	1	1
7	1	1	1	1	1	1.5	1	1	1	1
8	3	3	3	3	4	3	1	1.5	3.5	3
9	1	3	3	3	3	2	3	2	3	3
10	2	2	1	1	1	2	1	1	1	1
11	3	3	2	3	2	2.5	1.5	2	2	2
12	1	2	2	2	2	2.5	2	2	2	2
13	1	2	3	3	3	2.5	3	2	3	3
14	1	2	2	2	2	2.5	2	2	2	2
15	2	3	3	3	3	3	3	2.5	3	3
16	2	1	1.5	2	1	1	1	2.5	2	1.5
17	3	3	3	3	3	3	1.5	3	3	3
18	1	2	1	2.5	2	3	1	2	2	2
19	2	2.5	3.5	4	4	3.5	3	3	3	3
20	2	2	3	3	3	3	3	3	3	3
21	1	1	1	1	1	1	1	1	1	1
22	2	3	3.5	3.5	4	4	3	3	3.5	3.5
23	1	1	1	1.5	1	1	1	1	1	1
24	2	3	3	3.5	3	3.5	3.5	3	3.5	3
25	3	2	2.5	4	4	3.5	3	3	3	3
26	2	2	1.5	1.5	2	2	1.5	2	3	2
27	2	2	2	3	3	3.5	2.5	2	3	2.5
28	1	1	1	1	1	1	1	1	1	1
29	3	2	2	3	2	3	2	3	2	2
30	3	4	4	4	4	4	4	3	4	4
31	2	2	2	3	3	3	2	3	3	3
32	2	3	4	4	3	4	4	3	4	4
33	3	4	3	4	4	4	4	3	3.5	4
34	2	2	2	2	2	3	2	2	1	2
35	1	1	1	1	1	1	1	1	1	1
36	2	2.5	3	3	1	3	2.5	3	3	3
37	2	3	3	3	3	3.5	1	3	1	3
38	2	1	1	1	1	1	1	1	1	1
39	2	2	2	2	2	2	2	2	2	2
40	2	2	1	2	2	2	2	2	2	2
41	2	3	3	3	3	3	3	3	3	3
42	2	3	3	4	4	4	3	3.5	4	3.5
43	3	3	3	3	3	3.5	3	3	3	3
44	2	3	3	3.5	3.5	3.5	4	3.5	4	3.5
45	4	3	3	3	3	3.5	3	3	3	3
46	3	3	1.5	1	1.5	2.5	1	2.5	4	2.5
47	2	2	2	3	3	4	3	3	2	3
48	3	4	4	4	4	4	4	4	4	4
49	1	2	2	2	2	2	2	3	2	2
50	1	2	2	2	2	2	2	2	2	2
51	4	4	3	3.5	3	3.5	3	4	4	3.5
52	1	2	1	1.5	2	1	2	3	2	2
53	2	3	3	3	3	3	3	3.5	3	3
54	2	3.5	2	3	1	3	3	3	3	3
55	3	4	4	4	3	4	4	4	4	4
56	2	3	3	3	3.5	4	3	3	2	3
57	2	3	3.5	3.5	3.5	4	3	4	3	3.5
58	2	3	3	4	3.5	4	3	4	4	3.5

Round 2



Page	SME1	SME2	SME3	SME4	SME5	SME6	SME7	SME8	SME9	Median & Average
1	1	2	2	2	2	2	1	1	1	2
2	1	2	2	1	1	2	1	1	1	1
3	2	2	2	1	1	2	1	1	1	1
4	1	1.5	1	1	1	1	1	1.5	1	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1
8	2	3	2	1	3	4	4	1	1	2
9	1.5	3	2	1.5	2	3	1.5	2.5	1	2
10	1.5	2	1	1	1	1	1.5	1.5	1.5	1.5
11	1.5	3	2	1.5	2	2	2	1.5	1.5	2
12	2	2	2	1.5	2	2	2	2	1	2
13	2	2	3	2	2	3	2	2	2	2
14	2	2	2	1.5	2	2	2	2	1.5	2
15	3	3	3	3	3	1	2	2.5	2	3
16	2	1	2	2	2	1	2	2.5	1.5	2
17	2	3	3	2	2	3	2	2.5	2.5	2.5
18	2	2	2	2	2	2	2	2.5	2	2
19	2	2.5	3	3	3.5	4	1	3	1	3
20	3	2	3	2.5	3	3	2	3	3	3
21	1	1	1	1	1	1	1	1	1	1
22	2	3	3	2.5	3	3.5	3	2	3	3
23	2	1	1.5	1	1	1	1	1	1	1
24	3	3	3.5	2.5	3	3.5	3	3	3	3
25	2	2	3	2.5	2	2.5	3	3	4	2.5
26	2.5	2	1.5	2.5	3	1	2	3	3	2.5
27	2	2	2.5	2.5	3	3	2	3	2	2.5
28	1	1	1	2.5	1	1	1	1	1	1
29	2	2	2.5	2.5	3	2	2.5	3	2	2.5
30	3	4	3	3	4	4	3	3	4	3
31	3	2	3	3	3	3	3	3	1	3
32	3	3	3	4	4	4	3	3	2	3
33	3	4	3.5	3	4	3	1	3	3.5	3
34	3	2	2	3	2.5	2	2	2	1	2
35	2	1	2	1	1	1	1	1	1	1
36	3	2.5	2	3	3	2.5	3	3	3	3
37	4	3	3	3	3.5	4	2	2	4	3
38	2	1	1.5	2.5	1	1	1	2	1	1
39	3	2	2	3	3	2	2	3	2	2
40	2	2	2	2	2	2	2	1	1	2
41	3	3	3	3	3	3	3	3	3	3
42	3	3	3.5	3	4	4	4	3.5	3	3.5
43	3	3	3	3	3	3	3	3.5	3	3
44	4	3	3	3.5	4	4	3	3.5	3	3.5
45	3	3	3	3	3	3	3	3.5	3	3
46	3	3	2.5	3	3	1	2	3.5	3	3
47	3	2	2.5	3.5	3	3	2	3.5	1.5	3
48	4	4	3.5	3.5	4	4	3	3.5	4	4
49	3	2	2	3.5	2	2	2	2	2	2
50	3	2	2.5	3	2.5	2	2	3.5	2	2.5
51	4	4	3.5	3.5	4	4	4	3.5	4	4
52	3	2	2	3	3	1	2	3.5	2	2
53	3	3	3	3	3	3	3	4	2	3
54	3.5	3.5	2.5	3	3	3	3	3	3	3
55	3.5	4	4	4	4	4	3.5	4	4	4
56	3	3	3	4	3	3	3.5	4	2	3
57	4	3	3.5	4	3.5	3	4	3	3	3.5
58	4	3	3.5	4	4	2	4	4	4	4

Round 3



Page	SME1	SME2	SME3	SME4	SME5	SME6	SME7	SME8	SME9	Median & Average
1	1	2	2	2	2	2	1	1	1	2
2	1	1	2	1	1	2	1	1	1	1
3	2	2	2	1	1	2	1	1	1	1
4	1	1	1	1	1	1	1	1	1.5	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1
8	2	2	2	1	3	4	4	1	1	2
9	1.5	2	2	1.5	2	3	1.5	2.5	1	2
10	1.5	1	1	1	1	1	1.5	1.5	1.5	1
11	1.5	2	2	1.5	2	2	2	1.5	1.5	2
12	2	2	2	1.5	2	2	2	2	1	2
13	2	2	3	2	2	3	2	2	2	2
14	2	1	2	1.5	2	2	2	2	1.5	2
15	3	2	3	3	3	1	2	2.5	2	2.5
16	2	1	2	2	2	1	2	2.5	1.5	2
17	2	2	3	2	2	3	2	2.5	2.5	2
18	2	3	2	2	2	2	2	2.5	2	2
19	2	2	3	3	3.5	4	1	3	1	3
20	3	1	3	2.5	3	3	2	3	3	3
21	1	2	1	1	1	1	1	1	1	1
22	2	2	3	2.5	3	3.5	3	2	3	3
23	2	1	1.5	1	1	1	1	1	1	1
24	3	2	3.5	2.5	3	3.5	3	3	3	3
25	2	2	3	2.5	2	2.5	3	3	4	2.5
26	2.5	2	1.5	2.5	3	1	2	3	3	2.5
27	2	2	2.5	2.5	3	3	2	3	2	2.5
28	1	1	1	2.5	1	1	1	1	1	1
29	2	2	2.5	2.5	3	2	2.5	3	2	2.5
30	3	3	3	3	4	4	3	3	4	3
31	3	2	3	3	3	3	3	3	1	3
32	3	2	3	4	4	4	3	3	2	3
33	3	3	3.5	3	4	3	1	3	3.5	3
34	3	2	2	3	2.5	2	2	2	1	2
35	2	1	2	1	1	1	1	1	1	1
36	3	2	2	3	3	2.5	3	3	3	3
37	4	2	3	3	3.5	4	2	2	4	3
38	2	1	1.5	2.5	1	1	1	2	1	1
39	3	2	2	3	3	2	2	3	2	2
40	2	2	2	2	2	2	2	1	1	2
41	3	3	3	3	3	3	3	3	3	3
42	3	3	3.5	3	4	4	4	3.5	3	3.5
43	3	3	3	3	3	3	3	3.5	3	3
44	4	3	3	3.5	4	4	3	3.5	3	3.5
45	3	3	3	3	3	3	3	3.5	3	3
46	3	2	2.5	3	3	1	2	3.5	3	3
47	3	3	2.5	3.5	3	3	2	3.5	1.5	3
48	4	3	3.5	3.5	4	4	3	3.5	4	3.5
49	3	2	2	3.5	2	2	2	2	2	2
50	3	2	2.5	3	2.5	2	2	3.5	2	2.5
51	4	3	3.5	3.5	4	4	4	3.5	4	4
52	3	2	2	3	3	1	2	3.5	2	2
53	3	3	3	3	3	3	3	4	2	3
54	3.5	3	2.5	3	3	3	3	3	3	3
55	3.5	4	4	4	4	4	3.5	4	4	4
56	3	4	3	4	3	3	3.5	4	2	3
57	4	3	3.5	4	3.5	3	4	3	3	3.5
58	4	4	3.5	4	4	2	4	4	4	4

Conclusion

The methods and frameworks for making judgments about how to partition test takers into more than one category are numerous. As of the 1970's, there were "38 methods for setting criterion-referenced standards, and the number of methods has grown since then" (Cizek et al., 2001), and that reference was nearly 20 years ago.

One of the benefits of a typical bookmark or ID matching standard setting process is the ease of data entry for the facilitator because only the panelists' cutscores are entered into the facilitator's master spreadsheet. However, we recommend recording all of the panelists item-level ratings so that additional feedback can be provided to the panelists. It is easiest for the facilitator if the panelists record them in a spreadsheet that can be copied and pasted into the master spreadsheet, and this option also provides the most flexibility in terms of providing immediate feedback to the panelists on their individual ratings, such as in Figure 5. This is the option we recommend whenever possible. If it is not feasible to have panelists enter their ratings directly into a spreadsheet, their item-level ratings can still be captured by having the panelists write them on a rating sheet or on the ordered item booklet and having the facilitator or assistant hand-enter them into the master spreadsheet from the written source. In either case, group level feedback such as shown in Figure 3, 4 and 6 can be then shown to the panelists.

The enhanced feedback is intended to help facilitators guide the discussions between rounds of ratings and provide additional feedback to panelists to help them place their cutscores. The method described in this paper can also be used so that panelists do not need to choose individual cutscores and focus can instead be placed on the overall group-level results. The potential cut scores based on cumulative counts can be used instead of asking panelists to determine threshold regions, especially in Round 1. After the first feedback discussion, panelists may be ready to place their threshold regions if desired.

By recording item-level ratings and providing feedback as in Figure 3-6, the potentially frustrating task of determining panelist-level cutscores can be eased or eliminated, and instead panelists can focus their time on rating the items, discussing group-level feedback, refining their item-level ratings, and coming to consensus on the group-level results and recommendations. The meeting can run more smoothly, and the panelists may leave with less frustration and more willingness to engage in future test development work.

References

Cizek, G., & Bunch, M. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985918

Cizek, G., Raymond, M. R., & Reid, J. B. (2001). Who Made Thee a Judge? Selecting and Training Participants for Standard Setting. *American Registry of Radiologic Technologist*, 119–155.

Ferrara, S., Mcgraw-Hill, C. /, Perie, M., & Johnson, E. (2014). Matching the Judgmental Task with Standard Setting Panelist Expertise: The Item-Descriptor (ID) Matching Method. In *Journal of Applied Testing Technology* (Vol. 9, Issue 1). http://www.jattjournal.com/index.php/atp/article/view/48346

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. In *Educational Measurement: Issues and Practice* (Vol. 25, Issue 1, pp. 4–12). https://doi.org/10.1111/j.1745-3992.2006.00047.x

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

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/book/2008/gvcw

Authors

Dr. Deborah L. Schnipke is a Senior Psychometric Consultant at ACS Ventures, LLC and has over 20 years of experience working in measurement, providing psychometric expertise for all aspects of the test development process in a variety of fields. Deborah specializes in strategic planning and operational support for launching and sustaining assessment programs. She consults on the design and redesign of numerous testing programs, oversees test development operations for large- and small-scale programs, and conducts audits of testing programs. She is invested in ensuring that exams are reliable, valid, and fair, and in compliance with industry standards, such as the AERA/APA/NCME standards and NCCA and ISO 17024:2012 accreditation standards. She has experience as a speaker, reviewer, discussant, and author for major psychometric journals and conferences. Dr. Schnipke earned her Ph.D. in Quantitative Psychology from Johns Hopkins University.

Mr. Russell Keglovits is an Assessment Specialist who began working at ACS Ventures in June of 2019. Mr. Keglovits joined ACS after more than eight years of experience at the Nevada Department of Education and more than 10 years of experience as a high school Mathematics instructor. At the NV DOE, Russ was responsible for the authoring and operating of the accountability provisions expressed in state's federal education plan. This work included the collection and reporting of the state's education data. Russ has served on the eighth grade Mathematics Standing Committee for the *National Assessment of Educational Progress* (NAEP), has been a contributing member to several education advisory committees, and has led accountability standard setting projects. He has experience with assessment standard settings, validation studies, job task analysis, and item writing. He was also responsible for coordinating activities for the evaluation of many large-scale grant funded initiatives.