

# Standard Setting to Go<sup>i</sup>

Michael B. Bunch  
Measurement Incorporated

## Abstract

This paper describes a standard setting activity for a series of writing assessments for grades 3-10, conducted entirely online. Software development and deployment, management of training webinars and inter-round discussions, a vertical articulation webinar, and final evaluation of the process by panelists are described in detail. The paper concludes with lessons learned that may guide future online standard setting.

## Introduction

As other aspects of the educational assessment enterprise move online, it is logical that standard setting move there with them. The need for online standard setting is increasing as ownership of tests becomes more widely dispersed (as, for example, in the case of consortia or national tests). Gathering panelists in a single location for a district or state is daunting but doable. When those panelists are in multiple states (and even countries), the task becomes more daunting and considerably less doable, given the logistical challenges and expense associated with such an endeavor.

While there have been virtual standard settings in the recent past, they have been limited in scope. For example, in 2014, Smarter Balanced had over 2,000 individuals log in and place a single bookmark on a single test (Smarter Balanced Assessment Consortium, 2016). Those online panelists had no opportunity to discuss their bookmarks with others or even see the bookmarks others had set. For that same standard setting, nearly 500 additional panelists from 20 states met in a single location to discuss the tests, set bookmarks, and even see the bookmarks the online panelists had set.

But can standard setting be carried out entirely online with success? That was the question we set out to answer.

## Method

**Instruments.** This activity was conducted to set cut scores for the Writing Assessment Program (WrAP) published by the Educational Records Bureau (ERB). The WrAP is an essay-based assessment of student writing achievement. Students taking this test write a single essay on an assigned topic over a two-day period, with the first day devoted to a rough draft and the second day devoted to a final draft. Writing prompts address one of three genres: Narrative,

Informative, and Argument/Opinion. While WrAP includes both stimulus-based and non-stimulus-based prompts, this standard setting focused on non-stimulus-based prompts only.

Student essays are scored on six traits of writing achievement (each on a 6-point scale) to yield six subscores plus a total score. These scores are then transformed to scale scores and percentile ranks, with separate norms for independent, suburban, and international norming groups. In addition to the normative data provided in WrAP score reports, ERB plans to add criterion-referenced interpretations through the establishment of well-defined achievement levels. To this end, ERB had commissioned the development of achievement level descriptors (ALDs) based on learning progressions (LPs) that closely match the scoring rubrics for the various grade-band tests. Those ALDs were used for this standard setting

**Participants.** ERB recruited 43 educators (38 teachers and 5 administrators) from 19 (New York to California) states to recommend cut scores for selected grade/genre combinations of the WrAP. These 43 educators were organized into four panels. Table 1 shows the composition of each panel and its assignments.

**Table 1**  
**Panel Composition and Assignments**

Panel	Panelists	Packets	Total Essays
Grades 3-4	11	1 Narrative	35
Grades 5-6	9	1 Narrative; 1 Informative	35
Grades 7-8	12	1 Informative; 1 Argument/Opinion	35
Grades 9-10	11	1 Argument/Opinion	31

**Agenda.** Standard setting was carried out over a period of four consecutive days in July 2016. The agenda is summarized in Table 2.

**Table 2**  
**Agenda**

Day	Activity	Begin*	End*
1	Webinar: General Overview	11:00 AM	11:20 AM
	Overview of the tests and student responses	11:20 AM	12:20 PM
	Introduction to the Achievement Level Descriptors	12:20 PM	1:20 PM
	Break	1:20 PM	1:40 PM
	Introduction to the Body of Work Procedure	1:40 PM	2:25 PM
	Body of Work Practice Round	2:25 PM	2:55 PM
	Discussion of Practice Round	2:55 PM	3:25 PM

	Q&A/Wrap-Up	3:25 PM	3:40 PM
	Begin Round 1 (Range-finding)	3:40 PM	
2	End Round 1 (Range-finding)		3:40 PM
3	Webinar: Review of Round 1	11:00 AM	12:30 PM
	Round 2 (Pinpointing)	12:30 PM	Midnight
4	Webinar: Vertical Articulation	11:00 AM	12:30 PM
	Break	12:30 PM	1:00 PM
	Vertical Articulation	1:00 PM	3:30 PM

\* All times are Eastern Daylight

**Training.** Panelists received online training in a four-hour webinar that addressed the tests, the achievement level descriptors, and the Body of Work procedure. The traditional Body of Work procedure calls for panelists to assign each work sample to a single category, using the ALDs as their guide (cf. Cizek & Bunch, 2007, Chapter 9; Kingston & Tiemann, 2012). Cut scores are not actually calculated in the first (range-finding) round. Instead, that round is used to identify regions where cut scores may be found. During the second (pinpointing) round, panelists continue to assign each work sample to a category, and logistic regression is used to determine the cut score(s).

Wyse, Bunch, Deville, & Viger (2014) introduced a modification to this procedure. In their procedure, panelists are required only to identify the first work sample in the batch of score-ordered essays that seems to meet the criteria for a given level (i.e., the threshold value), again using the ALDs as their guides. Then, instead of using logistic regression to calculate cut scores, the investigator simply takes the median of the threshold values associated with the sample numbers entered by the panelists. Thus, for example, if a panelist were evaluating 35 work samples and trying to set two cut scores, he or she would enter 35 level designations using the traditional Body of Work approach or two sample numbers if using the Wyse *et al.* alternative.

All webinars were conducted with TurboMeeting. The initial training webinar began with brief introductions, followed by orientation to the tests and ALDs, presented via PowerPoint. During the presentation, there were multiple opportunities for panelists to ask questions, either orally or by using the Chat feature of the TurboMeeting software. After presentation of the tests and ALDs, panelists took a break.

After the break, panelists received instruction in the Body of Work procedure via PowerPoint, asked questions, and worked through a sample set of five essays to identify one threshold. After a lengthy discussion of the various locations panelists had chosen, the author answered their questions and terminated the session. Panelists then exited the webinar and logged in to accounts that had been set up for them to begin their first round of standard setting. Panelists had received and confirmed secure log-in information via e-mail prior to the training.

**Standard-setting software.** In 2014, Measurement Incorporated (MI) used software it had developed to carry out standard setting for the Smarter Balanced Assessment Consortium (Smarter Balanced). That software, Online Performance Level Setting (OPLS) was designed around a Bookmark procedure (cf. Cizek & Bunch, 2007). OPLS was modified to accommodate the Body of Work procedure for this activity.

*Software features for panelists.* Figures 1-11 show the basic features of the software from the panelists' perspective.

**Logging In**

**OPLS**

**Panelist Log in**  
Enter your User ID and Event Code to log in to your event.

User ID

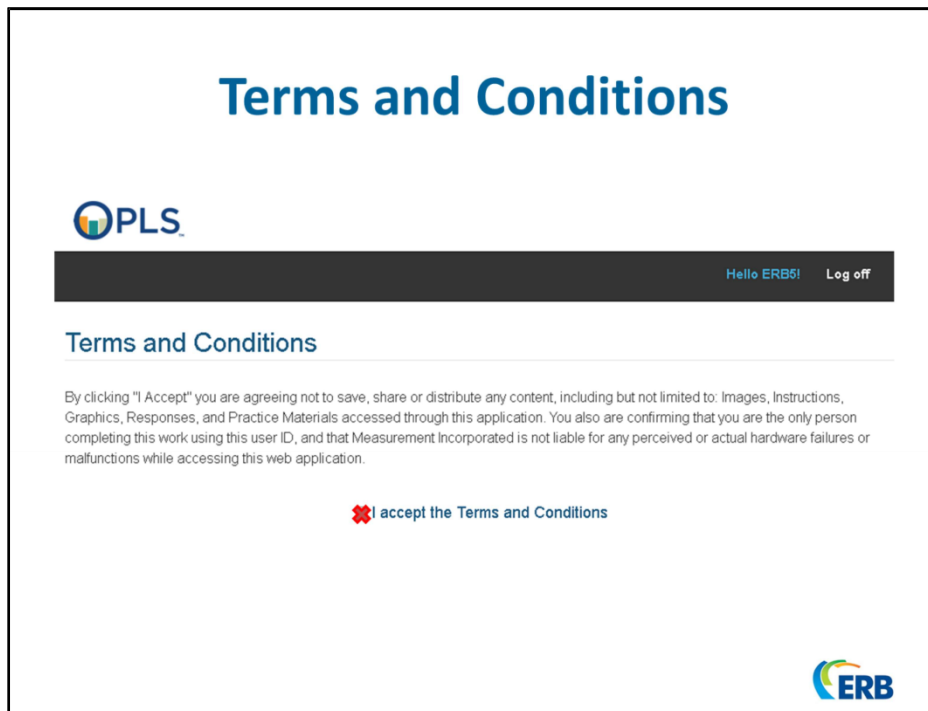
Event Code

Table   
Grades 5-6: Table 2  
Grades 7-8: Table 3  
Grades 9-10: Table 4

**ERB**

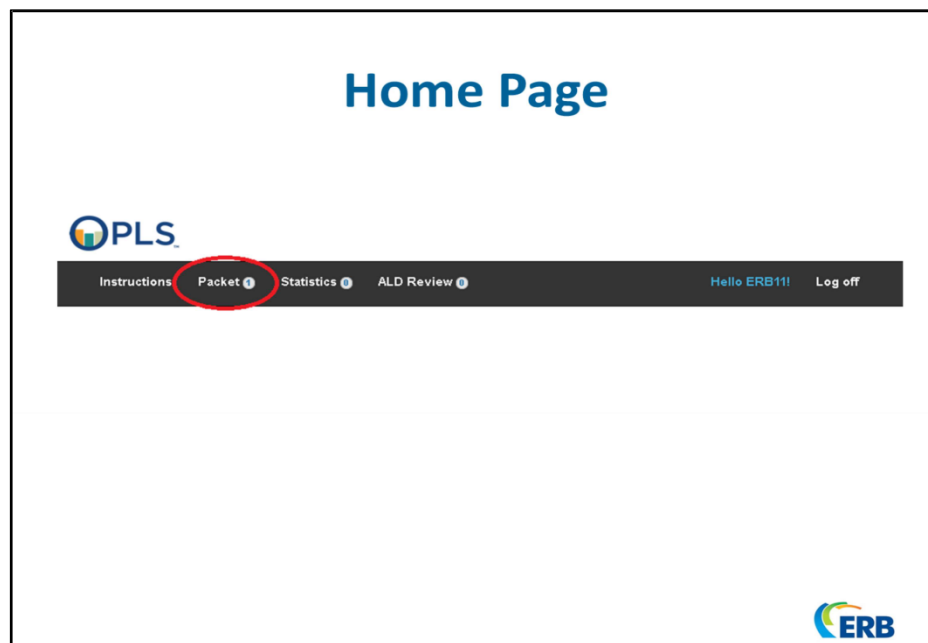
**Figure 1. Log-in screen**

Using log-in credentials previously sent by MI, panelists logged in to the system. Prior to gaining access, each panelist was required to click on and respond to a **Terms and Conditions** screen, shown in Figure 2.



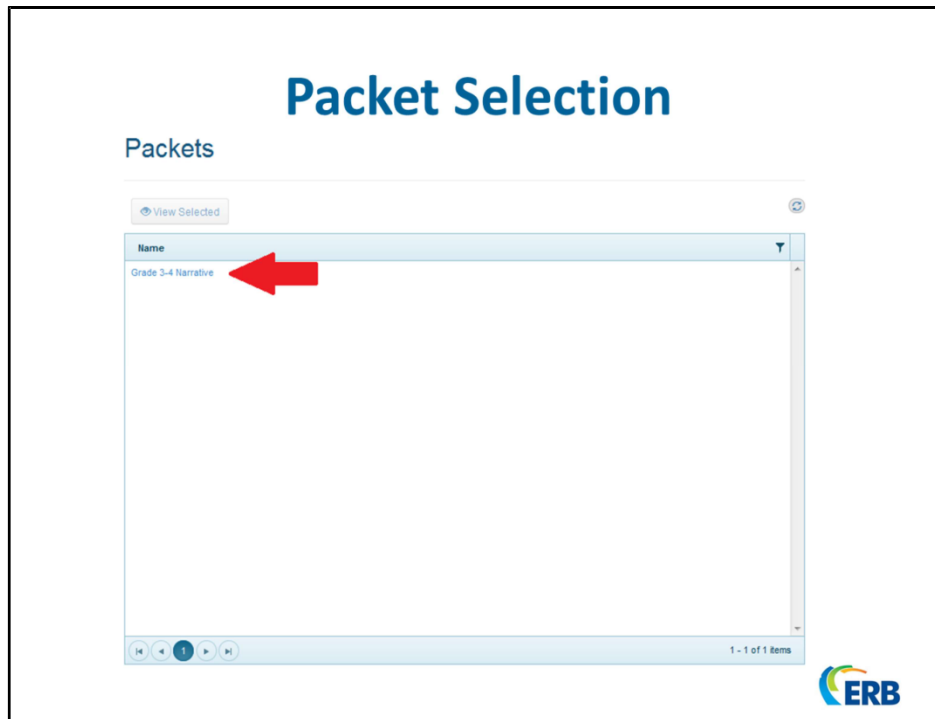
**Figure 2. Terms and Conditions screen**

After accepting the terms and conditions, panelists were led to the **Home Page**, where they selected the packets they were to review, as shown in Figure 3.



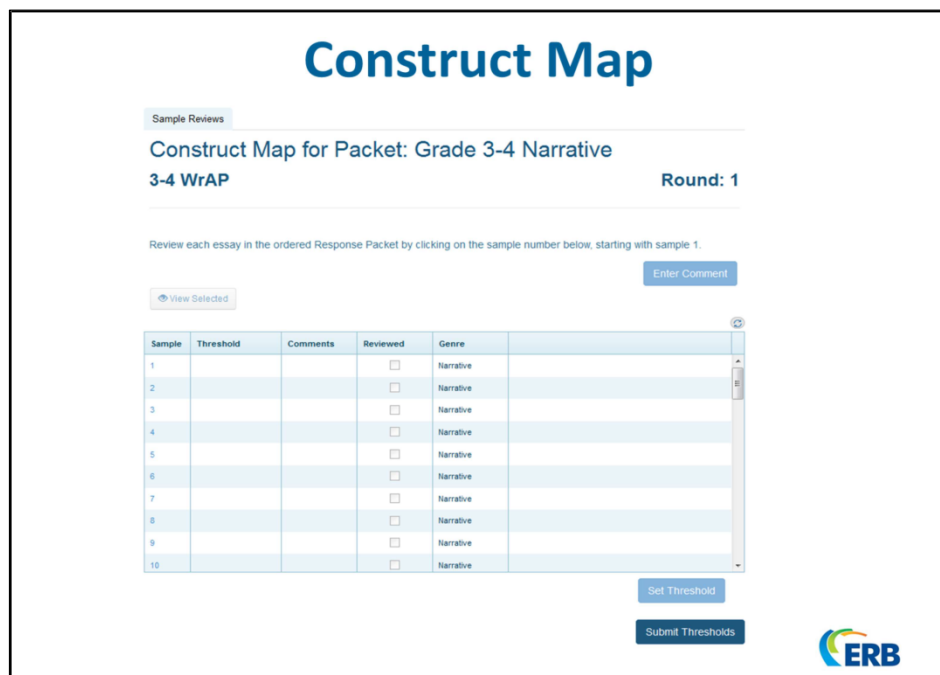
**Figure 3. Home Page screen**

Upon arriving on the **Home Page**, panelists would then select a packet of essays to review by clicking on the **Packet Selection** tab.



**Figure 4. Packet Selection screen**

Once they had selected a packet, the screen shown in Figure 5 appeared. The construct map shows for each essay to be rated its position in the ordered packet and its genre. Panelists selected essays to review by clicking on the essay number in the first column.



**5. Construct Map screen**

Clicking on any essay number in the first column would bring up the corresponding essay, as shown in Figure 6. Note that the essay in Figure 6 is Sample 1, the first sample in a packet ordered by score point. Subsequent essays will be of increasingly higher quality.

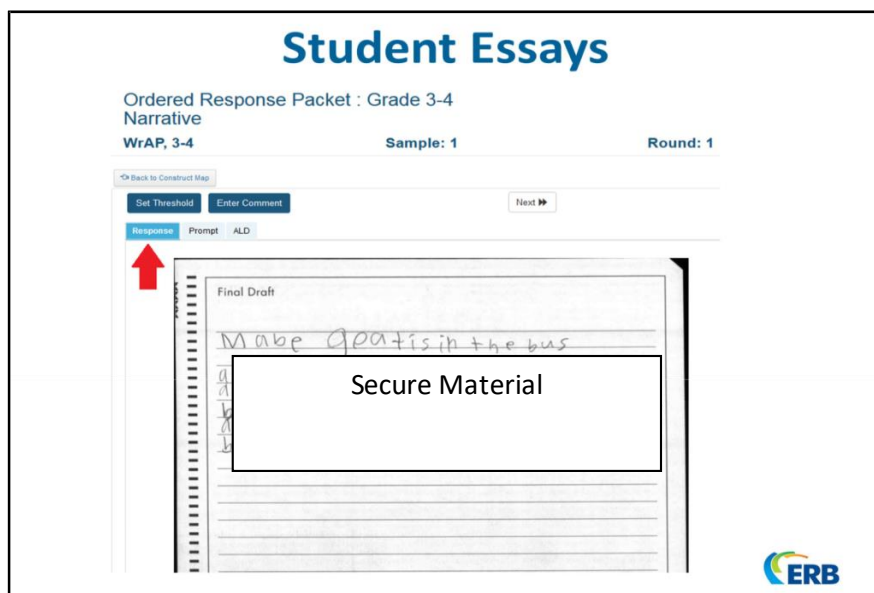


Figure 6. Student Essays screen

From the Student Essays screen, it was possible to navigate to the prompt or ALD. It was also possible to enter a comment or set a threshold by clicking the appropriate tab. Figure 7 shows the screen that would appear if the panelist had clicked **ALD**.

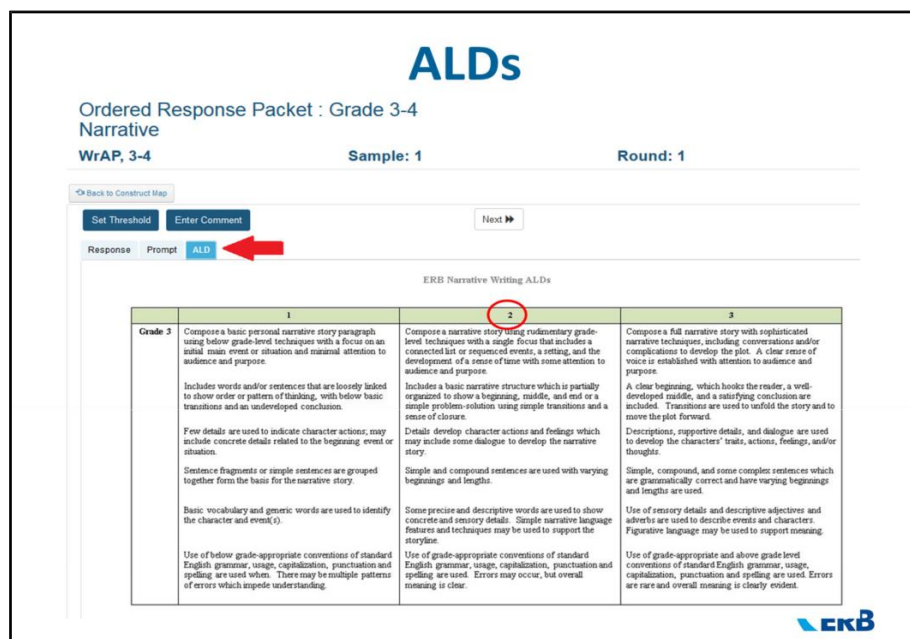
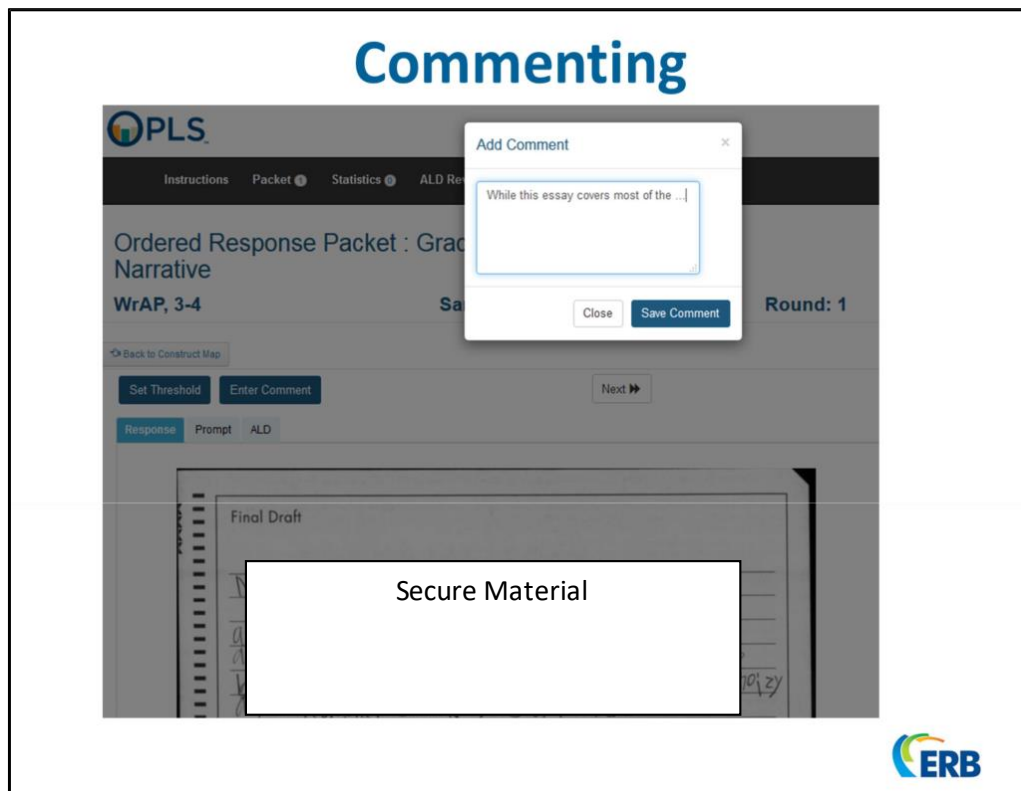


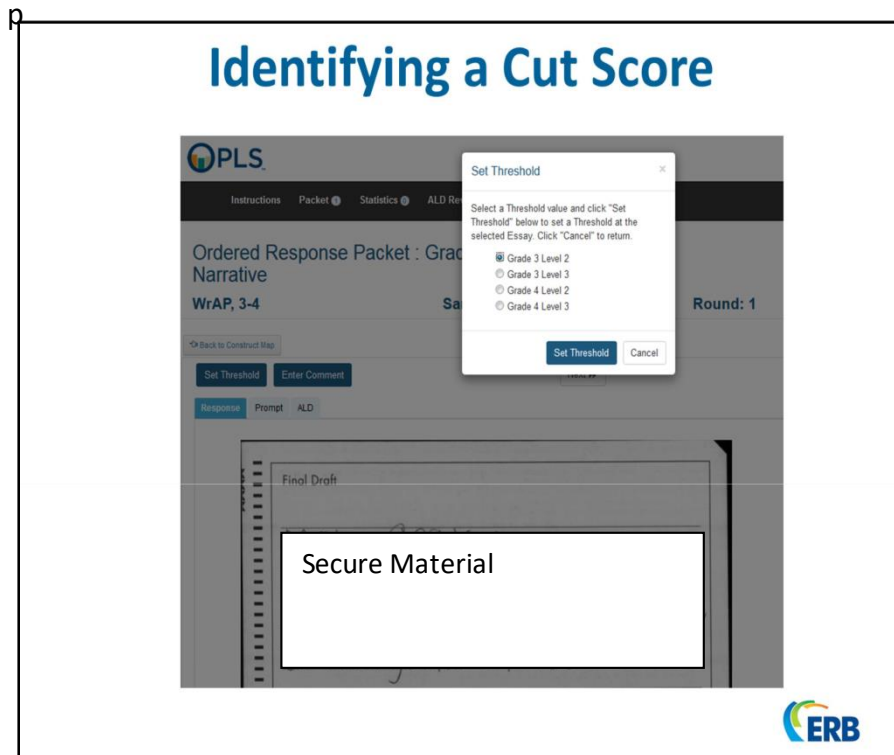
Figure 7. ALDs page

If the panelist had wanted to enter a comment, he or she would have clicked the **Enter Comment** tab and been taken to the screen shown in Figure 8.



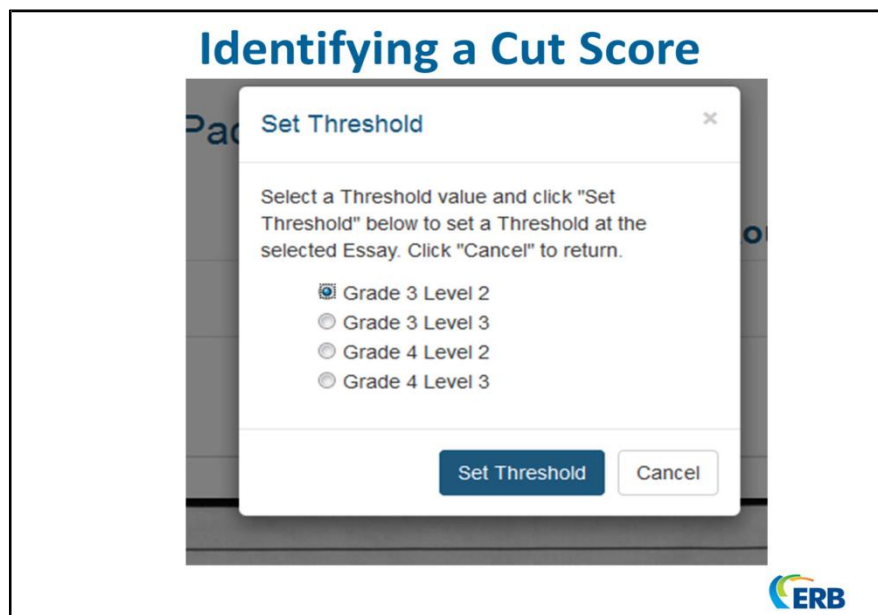
**Figure 8. Commenting screen**

At some point, a panelist would reach an essay in the ordered packet that seemed to meet the minimum criteria for Level 2. That panelist would then click on **Set Threshold** and go to the screen depicted in Figure 9. Again, keep in mind that the essay shown in Figure 9 is the same one that was in Figure 6; i.e., the first sample in the packet. It is unlikely that any panelist would set a threshold on the first page. This essay is simply used for illustrative purposes and to keep from exposing more essays than necessary.



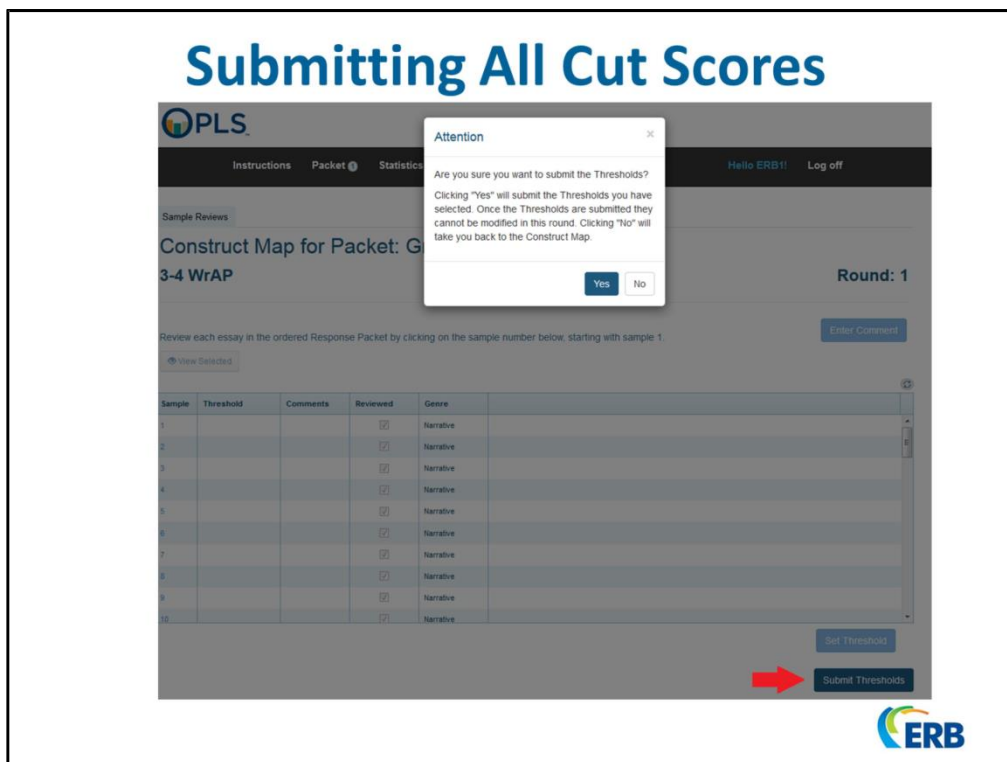
**Figure 9. Identifying a Cut Score screen**

Figure 10 shows a blow-up of the screen for **Identifying a Cut Score**.



**Figure 10. Blow-up of Identifying a Cut Score screen**

After locating the threshold essay for Level 2, panelists began looking for the threshold essay for Level 3, following the same procedures as they had in locating the Level 2 threshold. When they located that essay, they clicked the **Set Threshold** tab and entered the threshold. After identifying all thresholds, panelists were able to submit them and log out by returning to the **Construct Map** and clicking **Submit Thresholds** at the bottom of the screen. Figure 11 shows the screen that would appear when they did so.



**Figure 11. Submitting All Cut Scores screen**

Upon arriving at this screen, panelists would have an opportunity to go back and check their entries or submit them and log out. Upon submitting their thresholds and logging out, panelists were finished until the time scheduled for review of Round 1.

*Software features for facilitators.* Facilitators created the ordered packets and monitored panelist progress. Figures 12-14 show the software from the perspective of the facilitators.

As noted above, MI staff loaded hundreds of scored student essays into the software for use in Rounds 1 and 2. Facilitators could select packets for Round 1 and Round 2 simply by turning essay numbers on or off, as shown in Figure 12. One of the most arduous and time-consuming tasks of the Body of Work method is the elimination of certain Round 1 work samples and insertion of new work samples with scores in the threshold region, all in the short time between Round 1 and Round 2. Being able to complete this task with a series of clicks is a great leap forward.

## Assign Round Construct Maps

Event: Reload

Packet: Grade 5-6 Informative

Round: 1

[Back to List](#)

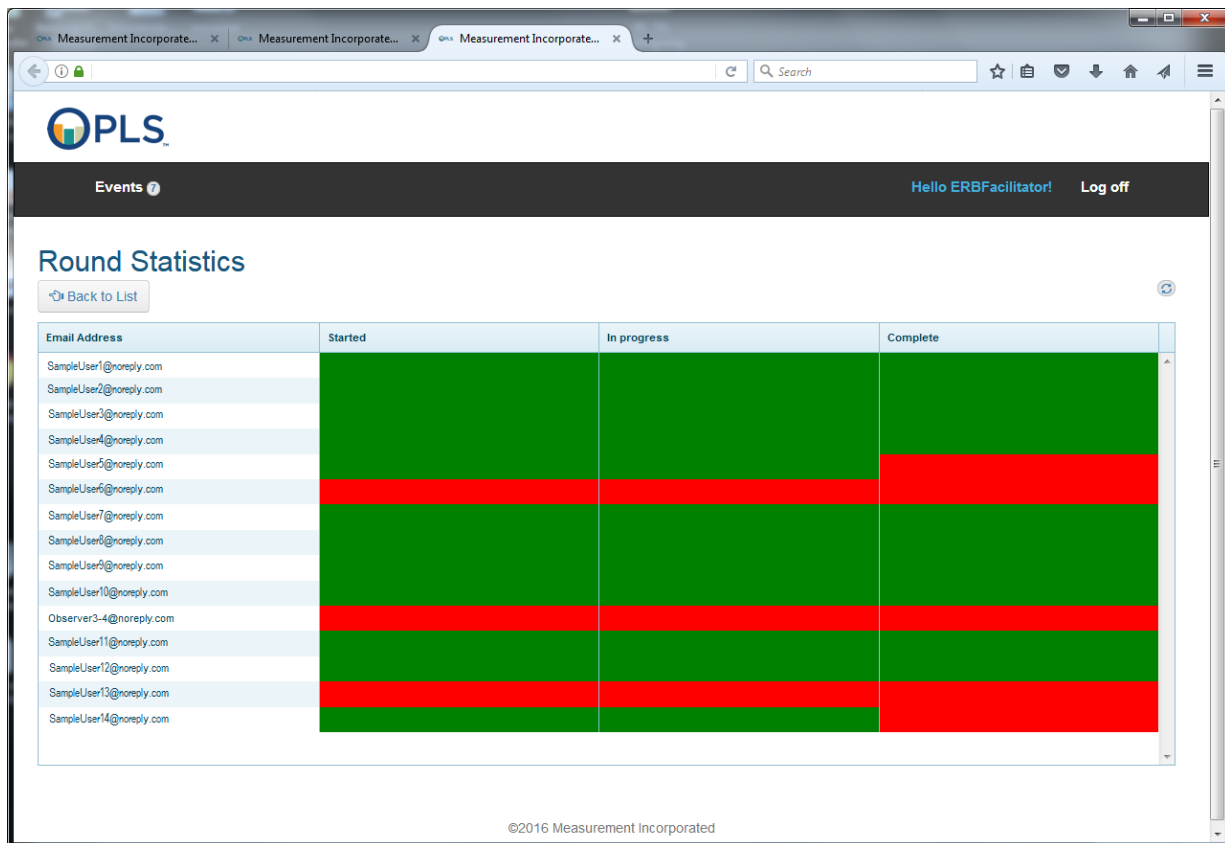
[Preview Construct Maps](#)

☒ Show only selected

Selected	Sample Number	Score	Genre
<input checked="" type="checkbox"/>	502	6	Informative
<input checked="" type="checkbox"/>	503	6	Informative
<input checked="" type="checkbox"/>	504	6	Informative
<input checked="" type="checkbox"/>	52	7	Informative
<input checked="" type="checkbox"/>	507	12	Informative
<input checked="" type="checkbox"/>	509	12	Informative
<input checked="" type="checkbox"/>	510	12	Informative
<input checked="" type="checkbox"/>	511	12	Informative
<input checked="" type="checkbox"/>	517	12	Informative
<input checked="" type="checkbox"/>	518	12	Informative
<input checked="" type="checkbox"/>	519	18	Informative
<input checked="" type="checkbox"/>	520	18	Informative
<input checked="" type="checkbox"/>	521	18	Informative
<input checked="" type="checkbox"/>	523	18	Informative

**Figure 12. Assign Round Construct Maps screen**

The software allowed facilitators to monitor in real time each panelist's movement through the packets, noting their activity and completion. At the same time, facilitators were available to all panelists in their groups via e-mail. Procedures for asking questions via e-mail were thoroughly addressed in the training and in written directions sent to panelists. Once panelists logged in, facilitators were able to track their progress from their facilitator login. In Figure 13, the green bars on the Round Statistics screen show tasks completed, and red bars show tasks not completed. A red bar all the way across the screen could indicate either an observer (who was not supposed to enter any cut scores) or a now-show panelist.



**Figure 13. Round Statistics screen**

Because each round had specific time limits (23 hours for Round 1 and 12 hours for Round 2, more than adequate time to complete the task), facilitators were able to control access to the packets by turning a round on or off, as shown in Figure 14.

## Event: ERB Standard Setting

### Edit Packet: Grade 5-6 Informative

[Back to List](#) [Save](#)

Rounds **2**
Statistics

Close Current Round
Start Next Round
View Round Statistics
Reopen Last Round

Name	Is Current	Round Start Date	Round End Date	Percent Compl...	Percent Started	Before Round ...	End Of Round Q...	Round Constr...	
<b>Status: Completed</b>									
1	False	07/18/2016 2:56 PM	07/20/2016 12:30 PM	64.3	78.6			2058, 2061, 2064, 2068, 2071, 2073, 2077, 2079, 2080, 2081, 2083, 2087, 2090, 2092, 2094,	Assign Round Construct Maps
<b>Status: Current</b>									
2	True	07/20/2016 12:30 PM		64.3	71.4			2073, 2075, 2077, 2079, 2080, 2535, 2081, 2083, 2084, 2085, 2087, 2090, 2544, 2091, 2092, 2094, 2095	Assign Round Construct Maps
<b>Status: Not Started</b>									
Ratings	False			0	0				Assign Round Construct Maps

0
No items to display

**Figure 14. Facilitator round control screen**

Facilitators and the author worked with the programmers to make sure all features were fully functional and user friendly. All facilitators participated in thorough training well in advance of the standard setting to make sure they would be able not only to navigate for themselves but offer assistance to panelists as well.

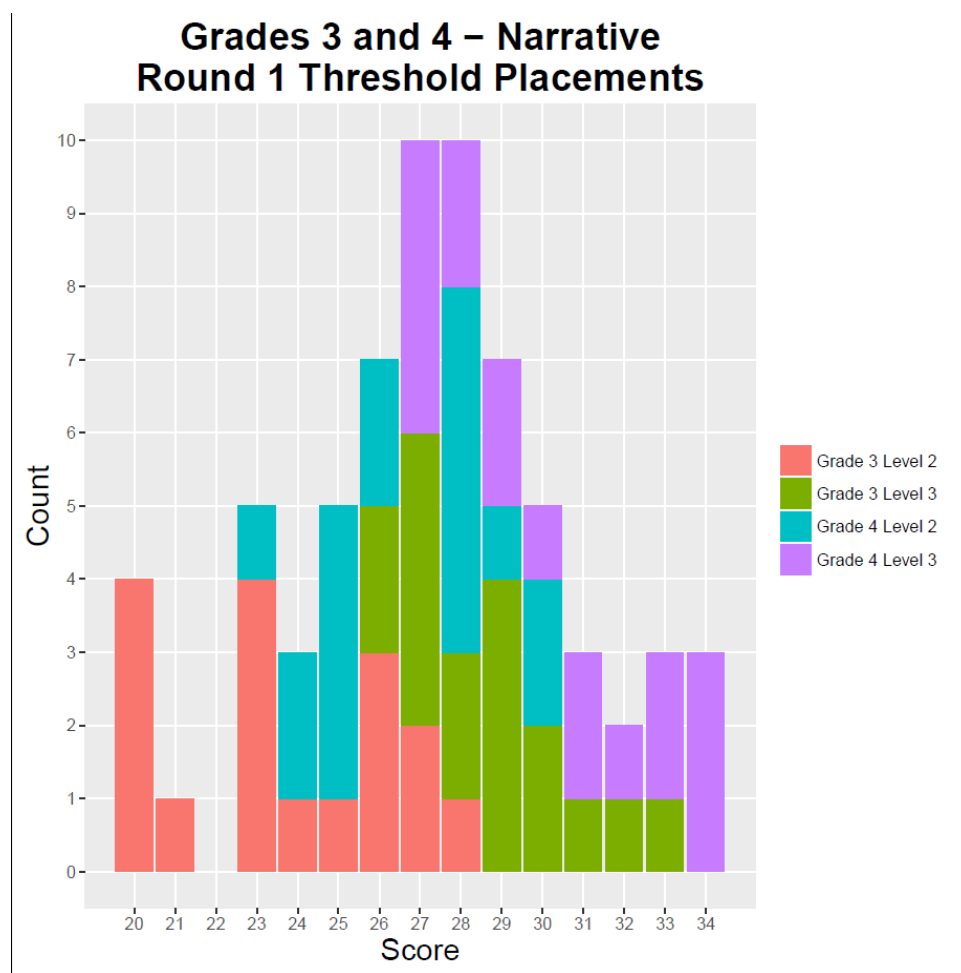
**Round 1.** Panelists logged in to OPLS and began the process of identifying thresholds. Each panelist had a specific set of essays to evaluate (as noted in Table 1). Panelists were instructed to start with the first essay in the first packet (Narrative for Grades 5-6 and Informative for Grades 7-8) and identify the first essay that met the requirements of the ALD for Level 2 for the lower grade.

After locating and marking that essay, they were to look for the first essay in the packet that met the requirements of Level 3 for that same grade. Having completed those two tasks, they were then to do the same thing for the higher grade. Panelists were instructed not to start over with the first essay the second time through the packet but to begin in the region where they had set the lower grade Level 2 threshold, reasoning that the higher grade threshold for Level 2 would not in all likelihood be below that of the lower grade. They received similar instructions for identifying the Level 3 threshold for the higher grade.

All panelists completed the task within the allotted time. MI staff collected the Round 1 data and computed distributions of thresholds as well as median cut scores. They prepared feedback

to share with panelists at the beginning of Round 2. Facilitators also identified cut score regions, deleted essays with scores outside those regions, and inserted other essays with scores in the regions identified by panelists in Round 1 as possible locations of cut scores. Because hundreds of essays had been loaded into the software before standard setting began, facilitators were able to view all essays for a given grade/genre and simply turn some on and others off between rounds by clicking check boxes next to each essay number. Because all essays had been entered into the software in score-point order, the essays selected for Round 2 were already in score-point order.

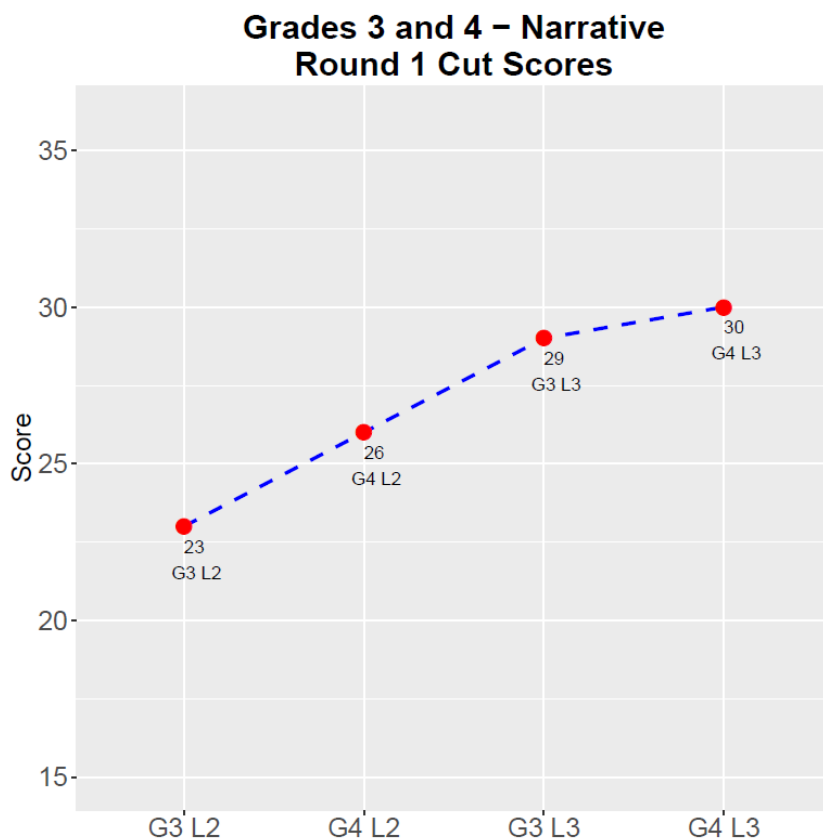
**Round 2.** MI conducted four separate webinars the morning of July 20. Each of the four facilitators had the results for Round 1 and shared them with panelists via TurboMeeting. The webinars began with a discussion of the panelists' experiences in Round 1, and the facilitators answered questions and encouraged panelists to describe their experiences for one another and share insights. The focus then shifted to the distribution of threshold placements. Figure 15 shows sample feedback regarding threshold distributions.



**Figure 15. Sample Round 1 threshold distribution feedback**

The purpose of sharing information such as this was to get panelists to tell why they had set a given threshold in a particular location and to listen to the rationales provided by other panelists. In Round 1, there is typically quite a bit of disagreement among panelists who have diligently applied the ALDs to the same set of essays as to where thresholds should be set. For example, in Figure 15, we can see that the essay that had a score of 27 was identified by two panelists as the threshold for Grade 3 level 2, by four panelists as the threshold for Grade 3 Level 3, and by four panelists as the threshold for Grade 4 Level 3. Moreover, some panelists had set the threshold for Grade 3 Level 3 lower than other panelists had set the threshold for Grade 3 Level 2. Discussions of these distributions consumed most of each of the 90-minute webinars.

After discussion of threshold distributions, facilitators presented median cut scores, based on panelists' Round 1 entries. Figure 16 shows a sample of that feedback. Panelists also received this information in tabular form.



**Figure 16. Round 1 feedback: Threshold medians**

At the end of each 90-minute webinar, panelists were again invited to log in to OPLS and complete Round 2 as they had completed Round 1. They had roughly 12 hours to complete this task. Once again, the facilitators monitored their progress. All panelists completed the task in the assigned time period.

**Vertical articulation.** Throughout Rounds 1 and 2, panelists had focused on specific essays that demonstrated the requirements of the ALDs for Levels 2 and 3 for a given grade and genre. The scores for these essays were reported in raw score terms (6-36 points). However, when comparing cut scores across grades, raw scores are an insufficient metric, as they are the same at every grade level. Instead, for comparison across grade bands, cut scores were converted to scale scores, which are not the same at each grade span but increase from span to span. Table 3 shows the range of scale scores by grade band.

**Table 3**  
**ERB WrAP Scale Score Range by Grade Band**

<b>Grade Span</b>	<b>Minimum Scale Score</b>	<b>Maximum Scale Score</b>	<b>Range</b>
Elementary (3-4)	176	547	371 points
Intermediate (5-6)	265	771	506 points
Middle School (7-8)	382	994	612 points
High School (9-10)	443	1479	1036 points
College Prep (11-12)	702	1548	846 points

On the afternoon of July 21, the author conducted another TurboMeeting webinar and presented the results of Round 2 to the vertical articulation committee (VAC). This committee consisted of three representatives from each of the four grade-span panels. The webinar began with an introduction to the concept of vertical articulation and the establishment of expectations. Normally, in a VAC, the focus is on percentages of students at each achievement level across grade levels (cf. Cizek & Bunch, 2007, Chapter 14). However, when a vertical scale is available, the focus can shift to examination of cut scores by grade level, relative to the vertical scale. Specifically, it is reasonable to expect cut scores to increase (in scale score terms) as one ascends the grade spans. For example, one would not likely expect the cut score for grade 9 to be lower than that for grade 8 on a vertical scale.

Next, the author described the procedures by which any modifications to cut scores would be made: formal motion, followed by second, followed by discussion, followed by vote. The author explained that since any motion to alter a cut score amounted to an override of a previous vote, it would require a 2/3 supermajority to pass. All understood and accepted this constraint.

## Results

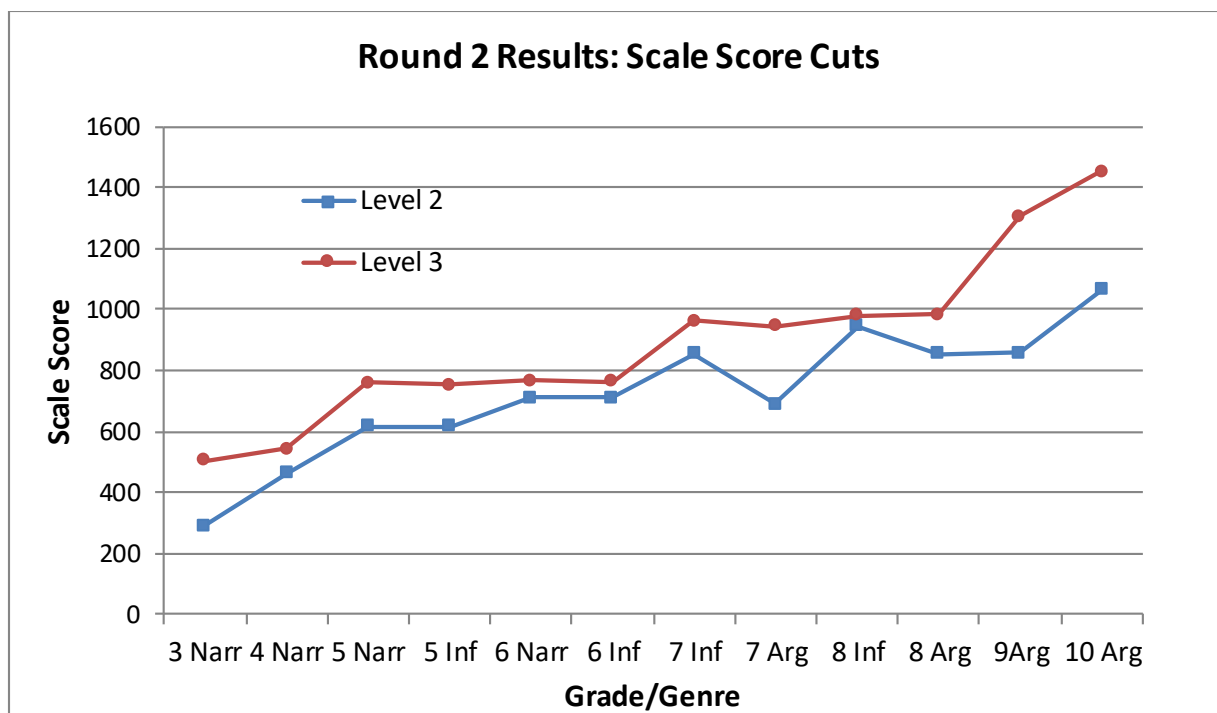
**Round 1.** As is typical in first rounds of standard setting, there was considerable variability with respect to the thresholds identified by panelists. On average, the range of cut scores was 10 points (10.4 points for Level 2 and 9.7 for Level 3). The greatest variability was for Grade 5 Informative Level 2 (15 points), Grade 7 Informative Level 2 (14 points), and Grade 7 Argument/Opinion Level 3 (17 points). These ranges were the subject of considerable discussion between Rounds 1 and 2.

**Round 2.** There was considerable movement away from extreme cut scores, even though all essays with scores identified in Round 1 as thresholds were included in Round 2. On average, the range of cut scores was 5.3 points for Level 2 and 6.0 points for Level 3, or 5.6 points overall. The reduction in range from Round 1 to Round 2 was significant: from 10.4 to 5.3 points for Level 2, from 9.7 to 6.0 for Level 3, and from 10.0 to 5.6 points overall. Particularly noteworthy are the three tests with large cut score ranges in Round 1:

- Grade 5 Informative Level 2 (15 points in Round 1 to 6 points in Round 2)
- Grade 7 Informative Level 2 (14 points in Round 1 to 2 points in Round 2)
- Grade 7 Argument/Opinion Level 3 (17 points in Round 1 to 4 points in Round 2)

These reductions in cut score range demonstrate the effects of inter-round discussion.

**Vertical articulation.** Results of Round 2 are shown graphically in Figure 17 in scale score terms. This is the graphic that was shown to VAC members during the opening webinar on the afternoon of July 21.



**Figure 17. Median cut scores after Round 2**

While the trend lines in Figure 17 are both generally ascending across grades, there is one dip in the Level 2 trend line. However, it is a within-grade drop. The Level 2 cut score for grade 7 Argument/Opinion was below that for grade 7 Informational. Previous discussions with panelists in both the 7-8 and 9-10 grade groups revealed that the Argument/Opinion genre is introduced in middle school and that students have difficulty with it until they become used to it. Consequently, they were willing to accept a lower cut score for Argument/Opinion at grades 7-8 than they were for Informational.

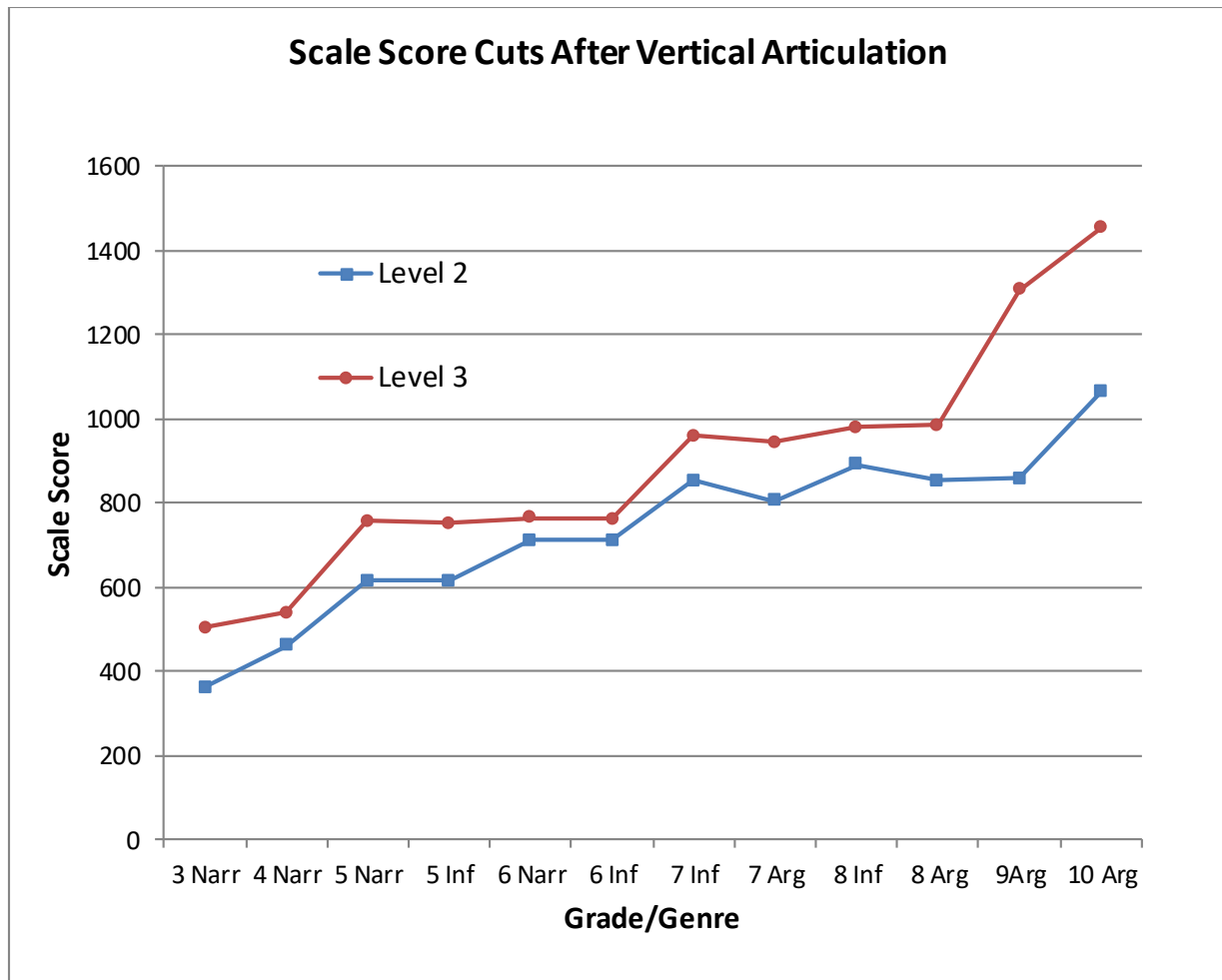
After orientation to the task, VAC members began to review the Round 2 cut scores and ask questions of one another and the author. The author then asked for motions from the floor to modify any specific cut score. Over the course of the next three hours, VAC members made three motions, each unanimously approved before taking up the next motion. In the end, there was a motion to approve all cut scores, unchanged originals as well as the three revised ones. Votes are recorded in Table 4. Final cut scores are shown in Table 5 and Figure 18. Cut scores that were changed by the VAC are highlighted in Table 5.

**Table 4**  
**Actions of the Vertical Action Committee**

<b>Motion</b>	<b>Made by</b>	<b>Seconded by</b>	<b>Vote</b>	<b>Action</b>
Change Grade 7 Argument/Opinion Level 2 cut from 21 to 23	Name Redacted	Name Redacted	Unanimous Approval	Passed
Change Grade 8 Informative Level 2 cut from 27 to 25	Name Redacted	Name Redacted	Unanimous Approval	Passed
Change Grade 3 Narrative Level 2 cut from 19 to 21	Name Redacted	Name Redacted	Unanimous Approval	Passed
Accept all cuts	Name Redacted	Name Redacted	Unanimous Approval	Passed

**Table 5**  
**Cut Scores After Vertical Articulation**

	<b>Articulated Raw Cut Scores</b>			<b>Articulated Scale Cut Scores</b>	
<b>Grade/ Genre</b>	<b>Level 2</b>	<b>Level 3</b>		<b>Level 2</b>	<b>Level 3</b>
3 Narrative	21	26		362	504
4 Narrative	24	31		462	541
5 Narrative	23	30		615	758
5 Informative	23	29		615	752
6 Narrative	26	32		711	766
6 Informative	26	31		711	763
7 Informative	24	28		853	961
7 Argument/Opinion	23	27		805	945
8 Informative	25	30		893	980
8 Argument/Opinion	24	31		853	985
9Argument/Opinion	20	25		858	1307
10 Argument/Opinion	22	30		1064	1454



**Figure 18. Final cut scores after vertical articulation**

At the end of all activities, panelists received an evaluation form via SurveyMonkey. They completed the survey and submitted their responses. Responses were then converted to numbers: Strongly Agree (SA) = 5; Agree (A) = 4; Undecided (U) = 3; Disagree (D) = 2; Strongly Disagree (SD) = 1. Results of the survey are shown in Table 6.

**Table 6**  
**Survey Results\***

<b>Statement</b>	<b>SD</b>	<b>D</b>	<b>U</b>	<b>A</b>	<b>SA</b>	<b>Mean</b>	<b>N</b>
<b>Opening Webinar and Round 1</b>							
The opening webinar contained information that was useful to me in understanding my task.	0	0	1	21	12	4.32	34
The information about the prompts and essays was helpful.	0	0	1	19	14	4.38	34
The information about the scoring rubrics was helpful.	0	0	4	19	11	4.21	34
The information about Achievement Level Descriptors was helpful.	0	0	4	16	14	4.29	34
The presentation on the Body of Work procedure was helpful.	0	0	3	18	13	4.29	34
The presenter helped me understand my Round 1 task.	0	1	0	14	19	4.50	34
The discussion among the panelists helped me understand my Round 1 task.	1	5	8	14	6	3.56	34
The chat feature was helpful.	0	3	2	12	16	4.24	33
The technical difficulties with the TurboMeeting made it difficult for me to complete my Round 1 task.	8	17	5	2	2	2.21	34
I left the webinar prepared to complete Round 1.	0	0	2	18	14	4.35	34
The additional information I received after the webinar was helpful to me in completing my Round 1 task.	0	1	5	21	7	4.00	34
<b>Wednesday Webinar and Round 2</b>							
The Wednesday webinar was well paced.	0	0	3	20	11	4.24	34
The information presented on screen was helpful to me.	0	1	3	20	10	4.15	34
The discussion about Round 1 threshold placements was helpful to me.	0	3	0	13	18	4.35	34
The presenter clearly explained differences between Round 1 and Round 2.	0	3	2	14	15	4.21	34
I left the Wednesday webinar prepared to complete Round 2.	0	2	2	14	16	4.29	34
I was able to get into the software to complete the task.	0	0	0	8	26	4.76	34
Once logged in, I was able to navigate the software without difficulty.	0	0	1	9	24	4.68	34
I am confident in the Round 2 thresholds that I set.	0	0	3	22	9	4.18	34

Statement	SD	D	U	A	SA	Mean	N
<b>Vertical Articulation</b>							
The introductory presentation helped me understand my task.	0	1	0	5	6	4.33	12
The presentation of data helped me understand and complete my task.	0	2	1	4	5	4.00	12
The comments of the other panelists helped me understand and complete my task.	0	0	1	5	6	4.42	12
Reviewing essays, ALDs, and prompts helped me make decisions.	0	1	1	5	5	4.17	12
I thought the decisions about changing cut scores were reached fairly.	0	0	0	6	6	4.50	12
I am satisfied with the final results that will go forward to ERB and the TAC.	0	0	0	6	6	4.50	12

\*(SD = Strongly Disagree; D = Agree; U = Undecided; A = Agree; SA = Strongly Agree)

Panelists were quite satisfied with the process overall. With regard to the opening webinar, mean scores on all but two statements were above 4.0 (Agree). The two exceptions were with regard to the discussion among panelists helping with the Round 1 task (Mean = 3.56, or about half way between Undecided and Agree). The statement about technical difficulties detracting from performance in Round 1 had a mean of 2.21 (Disagree), which is good because it demonstrates that even with the early audio difficulties, panelists were able to receive the instruction they needed and use it productively. By Round 2, panelists had become more familiar with the process and the software, as reflected in their responses to the evaluation statements.

The Vertical Articulation Committee responded very positively to all statements, with all 12 respondents either agreeing or strongly agreeing that they were satisfied with the final cut scores going forward to ERB. Other statements had mean responses ranging from 4.0 to 4.5, with 4.0 representing agreement and 5.0 representing strong agreement.

## Discussion

All told, panelists and facilitators spent close to ten hours engaged in virtual meetings and several more hours navigating ordered packets of essays online:

- Opening webinar – 4 hours
- Inter-round discussion – 1.5 hours
- Vertical articulation – 4 hours
- Packet review – up to 8 hours in two rounds (not necessarily continuous)

With the exception of some audio difficulties during the opening webinar, all virtual meetings went very smoothly. All panelists received advance log-in credentials and explicit instructions during the opening webinar and the inter-round discussion and were able to follow them with minimal difficulty. A review of the e-mail traffic during Rounds 1 and 2 and logs kept by the facilitators indicated that there were very few navigation problems and that the few that did occur were quickly solved. The evaluations by the panelists showed a high level of satisfaction with the process, comparable to that typically obtained in a face-to-face standard setting. A few comments from the SurveyMonkey evaluation are provided below.

- The information on the first day was a bit redundant. I thought we could have accomplished our task in far less time than was allotted, but as it turns out, we needed that time because of all the technical difficulties. Also, I really enjoyed the discussion on Day 3 and wish we had had some time for that kind of discussion on Day 1. I apologize if I said too much during our discussion on Day 3.
- Thank you for the opportunity to participate in this task. I enjoyed it. I know that the technical difficulties must have been very frustrating from your end, but I didn't mind. It was nice to see how well the TurboMeeting could actually work on day two.
- A terrific process that was well outlined and managed. Thank you for the opportunity to participate!
- I really enjoyed the work and would be happy to do it again.
- I truly appreciated the opportunity to participate in this process and learned much from doing so.
- I found the process very interesting and am glad I participated.

The technical difficulties referenced in two of the comments had to do with audio problems, primarily feedback from an unknown source. We were able to track down the source midway through the webinar. In the process, however, it became necessary to have everyone log out, wait 10 minutes, and then log back in. In the interim, MI support staff made some equipment changes. The audio problems recurred intermittently for the remainder of the opening webinar but were minimal. The author reverted to muting all participants and using the Raise Hand function of TurboMeeting to recognize other speakers one at a time. When all participants were on mute, there were no feedback issues.

**Lessons learned.** There are lessons to be learned when things go wrong (painful) as well as when they go right (pleasant). We present examples of both here.

*Painful.* The audio problem provided one of the most important lessons learned. The number of participants in the opening webinar was relatively small: 43 panelists plus co-facilitators and ERB staff, 49 in all. With a larger group, the need to mute participants would be even greater. We found that we could still conduct conversations using the Chat and Raise Hand functions of TurboMeeting. We strongly recommend that these features be employed and that all participants know in advance that they will be used.

*Pleasant.* For the Smarter Balanced standard setting in 2014, we conducted a field test of all software, instructions, and procedures. That field test revealed a number of weaknesses in instructions and procedures, and we were able to correct them before the main event. There was no opportunity to have a similar field test for this standard setting. However, the facilitators and the author did role play the inter-round discussions and created a generic script for all facilitators to use. That script proved to be extremely useful.

The author also conducted a dry run of the TurboMeeting setup and PowerPoint presentations with the facilitators days in advance of the opening webinar and noted what went well and what did not. A subsequent debriefing led to some minor modifications in approach. Facilitators also participated in training on the software, focusing on their own roles as well as the roles of panelists. Lesson learned: advance preparation is the *sine qua non* of a successful standard setting, whether it is virtual or face to face.

**Closing comments.** Virtual standard setting is not only possible but quite viable. In terms of logistics alone, a face-to-face meeting for 49 people (panelists plus facilitators plus client staff), assuming the need to pay for transportation, lodging, meals, and other miscellaneous travel expenses, would exceed \$70,000. The logistics costs for this standard setting were limited to the cost of the TurboMeetings. All other costs (principally people's time) would be the same either way, except for the fact that the TurboMeeting dramatically reduces transit time. It takes the better part of a day to travel from one coast to another and another day to travel back. For a webinar, it takes about a minute to log in; logging out takes even less time. Considering that the participants in this standard setting were scattered across the country, virtual standard setting made a lot of sense. The fact that it actually worked rather well made it seem even more reasonable.

The activity reported on here is part of a general effort on the part of Educational Records Bureau to make WrAP score reports more instructionally meaningful. Additional evidence of this effort may be found on their website, particularly the WRIIT linked resources (<https://www.erblearn.org/services/wrap-overview#introducing-wriit>).

## References

- Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.
- Kingston, N., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The body of work method. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations*. New York: Routledge.
- Smarter Balanced Assessment Consortium (2016). *2013=14 Technical Report*. (Chapter 10)  
<https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf>
- Wyse, A. E., Bunch, M. B., Deville, C. & Viger, S. G. (2014). A body of work standard setting method with construct maps. *Educational and Psychological Measurement*, 74 (2) 236-262.

---

<sup>i</sup> Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX, April 27, 2017.