

Chapter 10 Achievement Level Setting .....	3
Background .....	3
Achievement Level Setting Process .....	3
The Bookmark Procedure.....	3
Software development.....	7
Table 1 Software Development Timeline.....	8
Figure 1. Home Page With One Instruction Bar Expanded. ....	9
Figure 2. List of Resources Accessible From Home Page. ....	9
Figure 3. Sample Item Map. ....	10
Figure 4. Sample OIB Page With Selected-Response Item.....	11
Figure 5. Item Information Page. ....	11
Figure 6. OIB Page For Constructed-Response Item. ....	12
Figure 7. OIB Page Showing Links to Performance Task, Sample Student Response, and Rubric. ....	13
Figure 8. Comment.....	14
Figure 9. Set Bookmark Dropdown Box in the Item Map. ....	15
Figure 10. Submitting Bookmarks. ....	15
Design and Implementation of the Online Panel .....	16
Table 2. Numbers of Online Panelists, by Role, Grade, and Subject .....	17
Table 3. Impact of Online Panel Bookmark Placements: Percent of Students At or Above Level 3.....	17
Design and Implementation of the In-Person Workshop .....	18
Table 4. In-Person Workshop Panelists by Subject and Grade .....	19
Table 5. High-Level Agenda for Each In-Person Workshop.....	19
Table 6. Results of Round 1 of Bookmark Placement (Entries are Median Page Numbers). ....	21
Table 7. Results of Round 2 of Bookmark Placement (Entries are Median Page Numbers). ....	22
Table 8. Results of Round 3 of Bookmark Placement (Entries are Median Page Numbers). ....	23
Table 9. Round 3 Cut Score Recommendations: Scale Score Cuts and % At or Above.....	23
Table 10. Round 3 Questionnaire Results: Confidence in Cut Scores Recommended (Discounting Blanks).....	24
Table 11. Summary of Round 3 Evaluation Responses (Discounting Blanks) .....	24
Figure 11. Cross-Grade Review Graphic.....	26
Table 12. Cross-Grade Review Results .....	27

Approval by Chiefs .....	28
Table 13. Final Cut Scores Approved By Chiefs, With Impact Data. ....	29
Achievement Level Descriptors.....	30
Threshold ALDs .....	30
Comparison of final cuts to recommended cuts.....	31
Table 14. Comparison of Final Cuts to Those Recommended by the Cross-Grade Review Committees.....	31
ALD review.....	31
Findings and recommendations. ....	32
Range ALDs .....	32
Reporting ALDs.....	32
Long Range Validity Agenda for Performance Level Cut Scores.....	32
Validation Studies with Internal Variables.....	33
Validation Studies with External Variables.....	34
Organization and Implementation of Studies .....	36
Table 15. Validation Study Implementation Timeline. ....	36
References .....	39

## Chapter 10 Achievement Level Setting

### Background

In a request for proposals (RFP) issued in October 2013, Smarter Balanced called for a contractor to provide services for a multi-phase standard setting process and to plan and execute a comprehensive Communication Plan. The standard setting plan was to be executed in five phases from conducting distributed standard setting to finalizing achievement level descriptors. The communication plan was to “proactively explain the rationale for setting common achievement standards tied to the Common Core State Standards, describe the standard-setting process in layman’s terms, and make the case for approval of the performance standards derived from the standard setting process” (RFP, p. 10).

The word “standard” is used in many in educational and assessment contexts with varied meanings. To clarify the process of choosing proficiency thresholds, the consortium uses the term “achievement level setting” (ALS) which is used throughout this chapter in reference to the specific consortium activity. The assessment research literature calls this process “standard setting”, which is used in reference to the general process.

This chapter describes in some detail key outcomes of the ALS project. Specifically, this chapter documents the application and subsequent revision of achievement level descriptors, the in-person standard-setting activity, and the approved cut scores.

### Achievement Level Setting Process

Achievement level setting is the culminating set of activities in a four-year enterprise to create, field test, and implement a set of rigorous computer-adaptive assessments closely aligned to the Common Core and to provide guidance to educators regarding the achievements of their students, with particular reference to college and career readiness. The goal of the process is to identify assessment scores that delineate levels of achievement described by achievement level descriptors. Smarter Balanced has adopted four levels of achievement. For each grade and subject, there are three threshold cuts: Level 1 and level 2; Level 2 and Level 3; Level 3 and level 4. The division between Levels 2 and 3 is used as the proficiency criterion in accountability.

The ALS process used two components, an online panel that allowed broad stakeholder participation and provided a wide data set, and a more traditional in-person workshop that provided focused judgment from a representative stakeholder panel. The in-person workshop included a final cross-grade review stage. The online panel and in-person workshop used a Bookmark procedure (Lewis, Mitzel, Mercado, & Schultz, 2012), while the vertical articulation (cross-grade review) employed a procedure described by Cizek & Bunch (2007, Chapter 14). Details of both procedures are described in the sections below.

#### The Bookmark Procedure

The Bookmark standard setting procedure (Lewis et al., 2012) is an item response theory-based item mapping procedure developed in 1996 in response to the need for a robust standard setting procedure for high-stakes assessments of mixed format. Since 1996, it has become the most widely used procedure for setting cut scores on statewide assessments and other high stakes educational

assessments. Its psychometric foundation is well documented (e.g., Cizek & Bunch, 2007), and its usefulness has been well established through adoption of cut scores produced by Bookmark-based standard-setting activities.

### *Creating ordered item booklets.*

The bookmark method relies on presenting panelists with sets of test items sorted by difficulty and representing test content. This item collection is called the ordered item booklet (OIB). An important consideration when creating an ordered item booklet is to ensure appropriate content coverage. Psychometricians and content specialists from MI worked together closely to construct content specifications that matched Smarter Balanced guidelines with respect to targets and claims used to inform item and test development. The OIBs contained at least 70 items pages and with content weighted according to the specifications. Each OIB contained an entire performance task, that is, a set of 5-6 items/tasks all related to a set of stimuli. In order to minimize the reading load of the panelists, the ELA booklets included reading passages with a minimum of three associated items.

Since item order is the basis for panelist judgment, statistical considerations are of primary importance when building the OIBs. Thus, the booklets contained items that had a wide range of difficulty across the score scale with items at generally equal intervals of difficulty. All OIB items exhibited acceptable classical statistics, and showed no differential functioning. Combining the content and statistical constraints decreased the number of items for selection, but the final OIBs were very representative of the specified test content.

All OIBs were reviewed by MI, CTB, and Smarter Balanced's content and measurement experts. The reviews resulted in the removal and insertion of several items within each grade-content area until Smarter Balanced staff gave their final approval.

In a typical Bookmark procedure, each item in an OIB is mapped to an underlying construct in terms of the amount of that construct the examinee must possess in order to have a reasonable chance of answering the item correctly (in the case of a selected-response item) or obtaining a given score point or higher (in the case of an open-ended item or performance task).

In the three-parameter logistic (3PL) model, the Bookmark procedure relies on the basic relationship between person ability ( $\theta$ ) and item difficulty ( $b$ ), discrimination ( $a$ ), and pseudo-guessing ( $c$ ), where the probability of answering a dichotomously scored item correctly ( $P$ ) can be expressed as shown in equation (11.1).

$$P_j(X=1|\theta) = c_j + (1 - c_j)/\{1+\exp[-1.7a_j(\theta - b_j)]\} \quad (11.1)$$

where  $P_j$  is the probability of answering correctly,  $\theta$  is the ability required,  $a_j$  is the item discrimination index,  $\exp$  is the exponential function, and  $b_j$  is the item difficulty index. The way that guessing is accounted for is critical to the mapping. For most bookmark procedures, the  $c$  (pseudo-guessing) parameter is set to zero, so that the response probability specified is associated with the likelihood of a student knowing the correct response without guessing, as shown in equation (11.2). For this project, the two-PL model (with  $c$  set to 0) was used.

$$P_j(X=1|\theta) = 1/\{1 + \exp[-1.7a_j(\theta - b_j)]\} \quad (11.2)$$

For items with two or more score points, the probability of achieving any score  $k$  point or better given student ability  $P_{jk}(\theta)$  in a 2-parameter logistic model can be expressed as shown in equation 11.3 from Mitzel, Lewis, Patz & Green (2001).

$m_j$

$$P_{jk}(\theta) = \exp(z_{jk}) / \sum_{i=1}^{m_j} \exp(z_{ji}), \quad (11.3)$$

where  $m_j$  is the number of score points or steps for item  $j$ , and  $z_{jk} = (k - 1)\alpha_j - \sum_{i=0}^{k-1} \gamma_{ji}$ ;  $\alpha_j$  is the discrimination index of item  $j$ ,  $k$  is the number of this score point or step, and  $\gamma_{ji}$  is the step value for item  $j$  at step  $i$ . Thus, the probability of scoring at step  $k$  is a joint function of examinee ability, item discrimination, and the likelihood of obtaining any of the  $k - 1$  other scores. In this formulation, the value for a score of 0 (step 0) is set equal to zero; i.e.,  $\gamma_{j0} = 0$  for all items.

In practice, item maps show each item ordered in terms of the underlying value of  $\theta$  required to answer dichotomously scored items correctly and the value of  $\theta$  required to obtain at least each score point for multi-point items. Such maps may also contain other data, such as content domain, or other item metadata. It is also possible to show validation data.

In the Bookmark procedure, panelists are typically asked to find an item that a certain percentage of examinees at a critical threshold will be able to answer correctly. The cut score is identified at the point in an ordered item booklet beyond which panelists can no longer say that the target group would have the specified likelihood of answering correctly. The choice of that percentage is critical not only to defining the group of examinees but to defining the threshold between adjacent ability groups. This percentage is commonly called the RP value. In practice, users of the Bookmark procedure have employed 50 percent, 60 percent, 67 percent, and other values. For this project, upon the advice of the Technical Advisory Committee (TAC), RP50 was used.

Solving equation (11.2) for  $\theta$  produces equation (11.4):

$$\theta = b_j + \ln(1/P_j - 1) / (-1.7a_j) \quad (11.4)$$

where  $\ln$  is the natural logarithm and other values are as defined above. For any value other than 50%, the value for  $\ln(1/P_j - 1)$  is nonzero. However, when  $P_j = .50$ , the value of  $\ln(1/P_j - 1)$  reduces to  $\ln(1)$ , which is 0, and the value of  $\theta$  reduces to the item difficulty  $b_j$ , and item discrimination plays no part in the determination of the threshold ability level. Solving equation 11.3 for  $\theta$  involves the simultaneous calculation of the probabilities of obtaining each score point or better and is described in detail in Cizek & Bunch (2007). Thus the OIBs used in the consortium's achievement level setting process were ordered on the  $b$  parameter.

### *Item mapping.*

Item mapping allows individual items to be located along the scale score continuum so that interpretations about what students know and can do at individual scale score points may be facilitated. Item mapping is a component in the process of setting performance standards in the Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996). Item mapping is based in item response theory, exploiting the use of a common scale to express item difficulty and examinee proficiency. It requires the human judgmental process because panelists must make a decision about the response probability (RP; the likelihood that a person answers the item correctly) in order to align an item with a specific score point.

In addition to purely psychometric information, item maps may also contain item metadata (content standard, depth of knowledge, etc.) and other information. For this project, the contractor developed item maps that contained the content standard to which each item was aligned, the depth of knowledge associated with that item, ability level (expressed in scale score units), and, for the grade 11 tests, a region corresponding to a projection of college and career scale score levels on the ACT Assessment.

### *External data.*

Some of the items in the OIBs for grades 4, 8, and 11 are not Smarter Balanced items but actually come from other tests such as the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA). These items were embedded in the spring 2014 field test to provide panelists with an external reference range to the performance of students on other tests. They were to be used as part of the internal ALS process, not as a broad indicator.

In addition, for both Math and ELA in grade 11, panelists could see an area of the item map where ACT benchmark scores were projected. These benchmarks are estimates of scores students need to attain on the ACT in orderable to be considered ready to enter credit-bearing coursework at the postsecondary level.

Facilitators presented and discussed the external data rather briefly. Because many factors differentiate the Smarter Balanced tests from these other assessments, the facilitators maintained the focus of the panelists on the Smarter Balanced ALDs, relevant claims and targets, and the items in the OIBs.

### *Typical application of the Bookmark procedure.*

In a typical application of the Bookmark procedure, panelists receive extensive training in the content standards, the achievement level descriptors, the test to be reviewed, and the Bookmark procedure itself. This training typically takes a day or more. Panelists are then organized into small groups of 5-6 and instructed to review the OIB and place one or more bookmarks in accordance with the training procedures. Each such small group is led by a panelist serving as a table leader. Several such small groups make up a panel of 15 or more panelists, led by a facilitator in addition to the several table leaders. The facilitator provides ongoing instruction and leads discussions between rounds of item review. There are typically two or three rounds of item review.

After training in the bookmark procedure, panelists typically complete a practice round, setting a single bookmark in a very short OIB (usually 6-9 pages) and discuss the experience among themselves with leadership by the facilitator. Once all panelists confirm that they understand the process and the task, they begin Round 1.

In Round 1, panelists review the items in the OIB with a series of questions in mind:

- 1. What do you know about a student who responds successfully to this item; that is, what skills must a student have in order to know the correct answer?*
- 2. What makes this item more difficult than preceding items?*
- 3. Would a student at the threshold have at least a 50% chance of earning this point?*
  - Yes: Move on to the next item.*
  - No: Place your bookmark here.*

Panelists then place a bookmark on the first page in the OIB where they believe the student at the threshold for that level would NOT have at least a 50% chance of answering correctly. They complete this task once for each cut score.

After Round 1, bookmarks are tallied and shared among panelists for a given table. Those five or six panelists compare their Round 1 bookmark placements, discuss their rationales and understandings

of the threshold student at each level, and review the procedures for placing bookmarks. After this discussion, they answer a brief questionnaire indicating readiness to begin Round 2.

In Round 2, panelists once again review the OIB, this time bypassing pages that clearly did not contribute to bookmark placement. They continue to discuss the contents of the items with others at their table but place their own bookmarks. Using the same set of guiding questions they used in Round 1, panelists place a single bookmark for each cut score.

After Round 2, bookmarks are tallied, and a median bookmark for each cut score is calculated. These results are shared with the entire panel, along with impact data – percentages of students who would be classified at each level as well as percentages classified at or above all but the lowest level. Panelists, led by their facilitator, discuss the bookmark distributions as well as the impact data. After the discussion, panelists complete a brief questionnaire indicating their readiness to begin Round 3.

In Round 3, panelists once again review the OIB as in Round 2, but with the knowledge of the impact of their bookmark placements. Each panelist enters a bookmark for each cut score and submits his or her final bookmarks. After receiving the final median bookmark placements and associated impact data, panelists complete a final questionnaire and evaluation form.

#### **Software development.**

MI staff consulted with Smarter Balanced staff to create a detailed development schedule defining essential tasks and timelines for the online standard-setting web site. Using the approved requirements documentation, MI developers designed the online application and finalized in-person application software, continuing to work closely with Smarter Balanced staff in accordance with the timeline shown in Table 1.

Table 1 Software Development Timeline.

Software Development Task/Deliverable	Begin	End
Gather requirements/modify application design	2/3/14	3/7/14
Develop online tool	3/10/14	4/25/14
QA application	4/28/14	5/16/14
Receive additional SBAC feedback	5/19/14	5/30/14
Implement changes/make updates	6/2/14	8/1/14
Deploy and field test application	8/4/14	8/15/14
Address issues	8/18/14	9/19/14
Demonstrate for Smarter Balanced	9/22/14	10/3/14
Go live	10/6/14	10/20/14

The basic elements of the system were the home page, item map, and ordered item booklet. The home page contained all instructions, links to external resources (e.g., the Smarter Balanced website to allow panelists to take practice tests), and links to internal resources (instructions on applying the Bookmark procedure, Common Core State Standards, and Achievement Level Descriptors). The item map had many features that could be turned on or off, depending on the round and nature of the task to be performed. The OIB contained the items as well as metadata, sample responses, and links to the ALDs.

***The home page.***

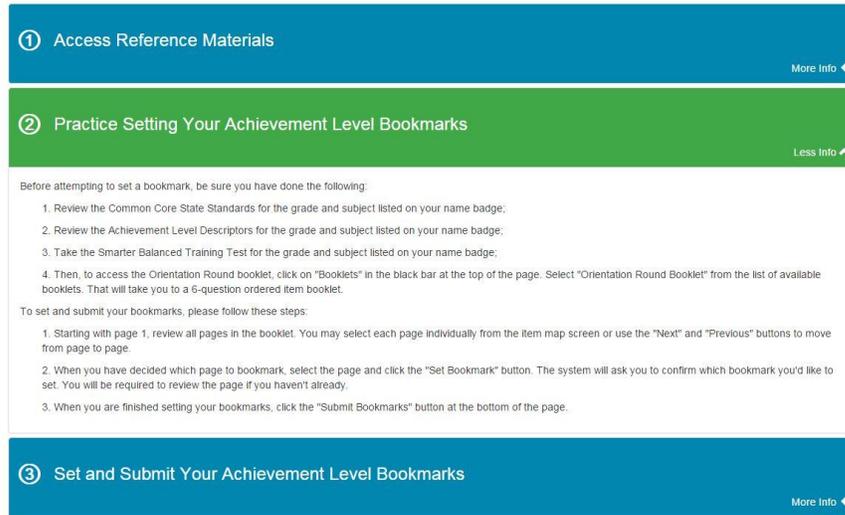
The home page contained all instructions plus links to additional resources. It consisted of four numbered, horizontal bars that could be expanded to reveal detailed information about each step of the process, as shown in Figure 1

Figure 1. Home Page With One Instruction Bar Expanded.

### How to Use the Tool

During the In-Person Panel for Achievement Level Setting, you will need to complete the three steps listed below. To learn more about a step, click on the "More Info" button next to each one.

At certain points in this process, the system will present you with a short questionnaire. You must enter a response to all questions, and submit the questionnaire, before moving on to the next task. You may save your questionnaire at any time by clicking the "Save" button.



**1 Access Reference Materials** More Info ◀

**2 Practice Setting Your Achievement Level Bookmarks** Less Info ▲

Before attempting to set a bookmark, be sure you have done the following:

1. Review the Common Core State Standards for the grade and subject listed on your name badge;
2. Review the Achievement Level Descriptors for the grade and subject listed on your name badge;
3. Take the Smarter Balanced Training Test for the grade and subject listed on your name badge;
4. Then, to access the Orientation Round booklet, click on "Booklets" in the black bar at the top of the page. Select "Orientation Round Booklet" from the list of available booklets. That will take you to a 6-question ordered item booklet.

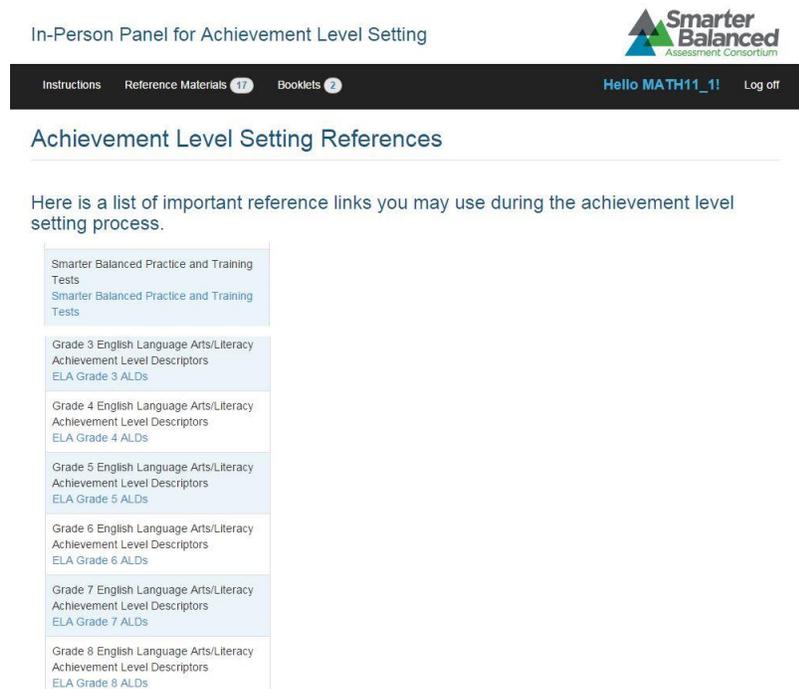
To set and submit your bookmarks, please follow these steps:

1. Starting with page 1, review all pages in the booklet. You may select each page individually from the item map screen or use the "Next" and "Previous" buttons to move from page to page.
2. When you have decided which page to bookmark, select the page and click the "Set Bookmark" button. The system will ask you to confirm which bookmark you'd like to set. You will be required to review the page if you haven't already.
3. When you are finished setting your bookmarks, click the "Submit Bookmarks" button at the bottom of the page.

**3 Set and Submit Your Achievement Level Bookmarks** More Info ◀

The home page contained a list of all resource materials, accessible through hyperlinks, as shown in Figure 2.

Figure 2. List of Resources Accessible From Home Page.



In-Person Panel for Achievement Level Setting 

Instructions Reference Materials 17 Booklets 2 Hello MATH11\_1! Log off

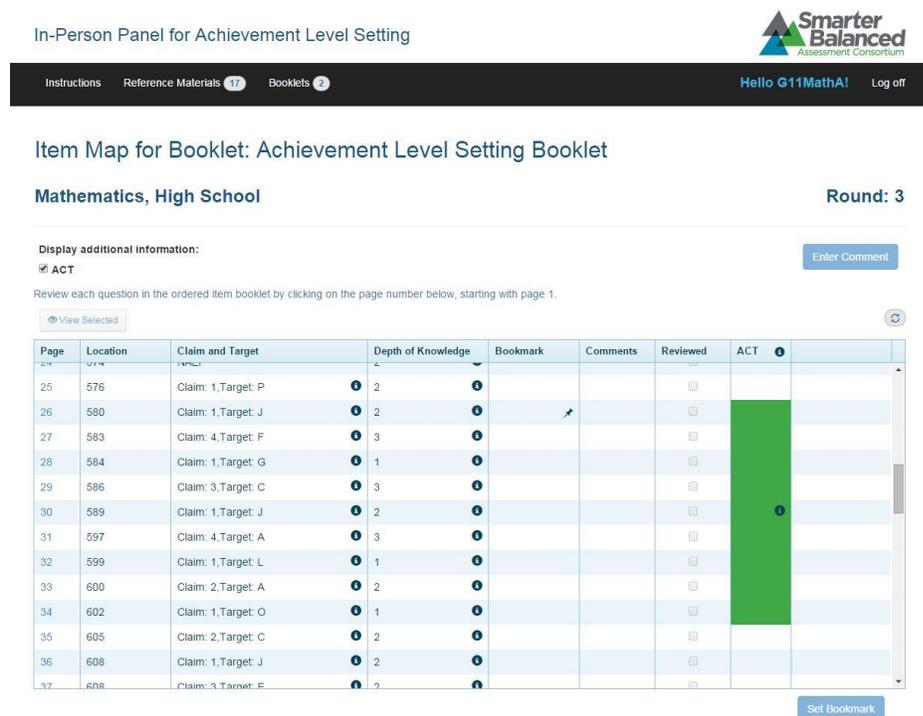
### Achievement Level Setting References

Here is a list of important reference links you may use during the achievement level setting process.

- Smarter Balanced Practice and Training Tests
- Smarter Balanced Practice and Training Tests
- Grade 3 English Language Arts/Literacy Achievement Level Descriptors  
ELA Grade 3 ALDs
- Grade 4 English Language Arts/Literacy Achievement Level Descriptors  
ELA Grade 4 ALDs
- Grade 5 English Language Arts/Literacy Achievement Level Descriptors  
ELA Grade 5 ALDs
- Grade 6 English Language Arts/Literacy Achievement Level Descriptors  
ELA Grade 6 ALDs
- Grade 7 English Language Arts/Literacy Achievement Level Descriptors  
ELA Grade 7 ALDs
- Grade 8 English Language Arts/Literacy Achievement Level Descriptors  
ELA Grade 8 ALDs

The online item map page (see Figure 3) allowed panelists to review their progress, navigate through the ordered item booklet pages, access relevant item data, and submit their bookmarks. The Item Map drop-down menu allowed panelists to select and view their current results as well as the results from their previous round. Hovering over a comment indicator displayed the comments they submitted for a specific item during a round.

Figure 3. Sample Item Map.



Page	Location	Claim and Target	Depth of Knowledge	Bookmark	Comments	Reviewed	ACT
25	576	Claim: 1, Target: P	2			<input type="checkbox"/>	
26	580	Claim: 1, Target: J	2			<input type="checkbox"/>	
27	583	Claim: 4, Target: F	3			<input type="checkbox"/>	
28	584	Claim: 1, Target: G	1			<input type="checkbox"/>	
29	586	Claim: 3, Target: C	3			<input type="checkbox"/>	
30	589	Claim: 1, Target: J	2			<input type="checkbox"/>	
31	597	Claim: 4, Target: A	3			<input type="checkbox"/>	
32	599	Claim: 1, Target: L	1			<input type="checkbox"/>	
33	600	Claim: 2, Target: A	2			<input type="checkbox"/>	
34	602	Claim: 1, Target: O	1			<input type="checkbox"/>	
35	605	Claim: 2, Target: C	2			<input type="checkbox"/>	
36	608	Claim: 1, Target: J	2			<input type="checkbox"/>	
37	608	Claim: 3, Target: F	2			<input type="checkbox"/>	

Each OIB page displayed item-specific information including a preview of the item, item statistics, answer key(s), and associated passages and scoring rubrics. Additionally, the OIB page was designed to allow the panelist to make a comment about an item and store that comment for later review. The OIB page included a link to the Achievement Level Descriptor (ALD) for each test. Figure 4 shows a sample selected-response item, while Figure 5 shows the associated item information page, and Figure 6 shows a page for a constructed-response item (in this case, a performance task).

The item map and OIB pages were designed to allow panelists to toggle back and forth. Panelists could gain access to any page in the OIB by clicking that page number in the item map and return to the item map by clicking “Back to Item Map” at the top or bottom of the page. Each OIB page displayed the item, item statistics, rubrics, passages, and sample responses. Additionally, the OIB page was designed to allow the panelist to specify a cut score or navigate to the next or previous OIB page.

All items presented in the OIB were in static, portable data file (pdf) format rather than in interactive format as they had been in the practice tests on the Smarter Balanced website or as administered in the spring 2014 field test. The decision to render items in a static format was based on concerns about the rendering of the interactive versions of items on an uncontrollable array of online panelist

devices and browsers. By displaying a static image or PDF of the item, it was possible to ensure that every panelist saw exactly the same rendering of the item for review independent of the platform used.

Figure 4. Sample OIB Page With Selected-Response Item.

Ordered Item Booklet: Achievement Level Setting Booklet

Mathematics, High School Page: 06 Round: 2

← Back to Item Map

Set Bookmark Enter Comment

← Previous Next →

Item Question Information Passages and Other Materials 0 Achievement Level Descriptors 1

This item permits calculator use.

**12155**

The formula for the rate at which water is flowing is  $R = \frac{V}{t}$ , where

- $R$  is the rate,
- $V$  is the volume of water measured in gallons ( $g$ ), and
- $t$  is the amount of time, in seconds ( $s$ ), for which the water was measured.

Select an appropriate measurement unit for the rate.

(A)  $gs$

(B)  $\frac{g}{s}$

(C)  $\frac{s}{g}$

(D)  $\frac{1}{sg}$

← Back to Item Map

Figure 5. Item Information Page.

In-Person Panel for Achievement Level Setting



Instructions Reference Materials 17 Booklets 2 Hello G11MathA! Log off

Ordered Item Booklet: Achievement Level Setting Booklet

Mathematics, High School Page: 31 Round: 2

← Back to Item Map

Set Bookmark Enter Comment

← Previous Next →

Item Question Information Passages and Other Materials 3 Achievement Level Descriptors 1

Page	31
Location	597
Claim and Target	Claim: 4, Target: A
Depth of Knowledge	3
Answer Key	See Passages and Other Material Tab

← Back to Item Map

Figure 6. OIB Page For Constructed-Response Item.

In-Person Panel for Achievement Level Setting 

---

Instructions Reference Materials **17** Booklets **2** Hello G11ELAA! Log off

---

Ordered Item Booklet: Achievement Level Setting Booklet

English Language Arts/Literacy, High School Page: 03 Round: 1

---

[Back to Item Map](#)

Set Bookmark Enter Comment

Previous Next

Item Question Information Passages and Other Materials **3** Achievement Level Descriptors **1**

---

**62019**

**Student Directions for Part 2**

You will now review your sources, take notes, and plan, draft, revise, and edit your article. You may use your notes and refer to the sources. Now read your assignment and the information about how your article will be scored; then begin your work.

**Your assignment:**  
 After completing your research, you share your findings with your teacher. She is impressed with your work. As a final project for your psychology class, everyone must write an article for the Psychology Club's website. Your teacher suggests writing about malleable intelligence, and you decide this is a good idea. The audience for your article will be other students, teachers, and parents.

Using more than one source, craft a thesis to explain the concept of malleable intelligence. Once you have a thesis, select the most relevant information to support your thesis. Then, write a multi-paragraph explanatory article explaining your thesis. Clearly organize your article and elaborate on your ideas. Develop your ideas clearly and use your own words, except when quoting directly from the sources. Be sure to reference the source title or number when quoting or paraphrasing details or facts from the sources.

**Explanatory Scoring**  
 Your explanatory article will be scored using the following:

- 1. Organization/purpose:** How well did you state your thesis, and maintain your thesis with a logical progression of ideas from beginning to end? How well did you narrow your thesis so you can develop and elaborate the conclusion? How well did you consistently use a variety of transitions? How effective was your introduction and your conclusion?
- 2. Elaboration/evidence:** How well did you integrate relevant and specific information from the sources? How effective were your elaborative techniques? How well did you clearly state ideas using precise language that is appropriate for your audience and purpose?
- 3. Conventions:** How well did you follow the rules of grammar usage, punctuation, capitalization and spelling?

**Now begin work on your article.** Manage your time carefully so that you can:

- plan your multi-paragraph article
- write your multi-paragraph article
- revise and edit the final draft of your multi-paragraph article

Word-processing tools and spell check are available to you.

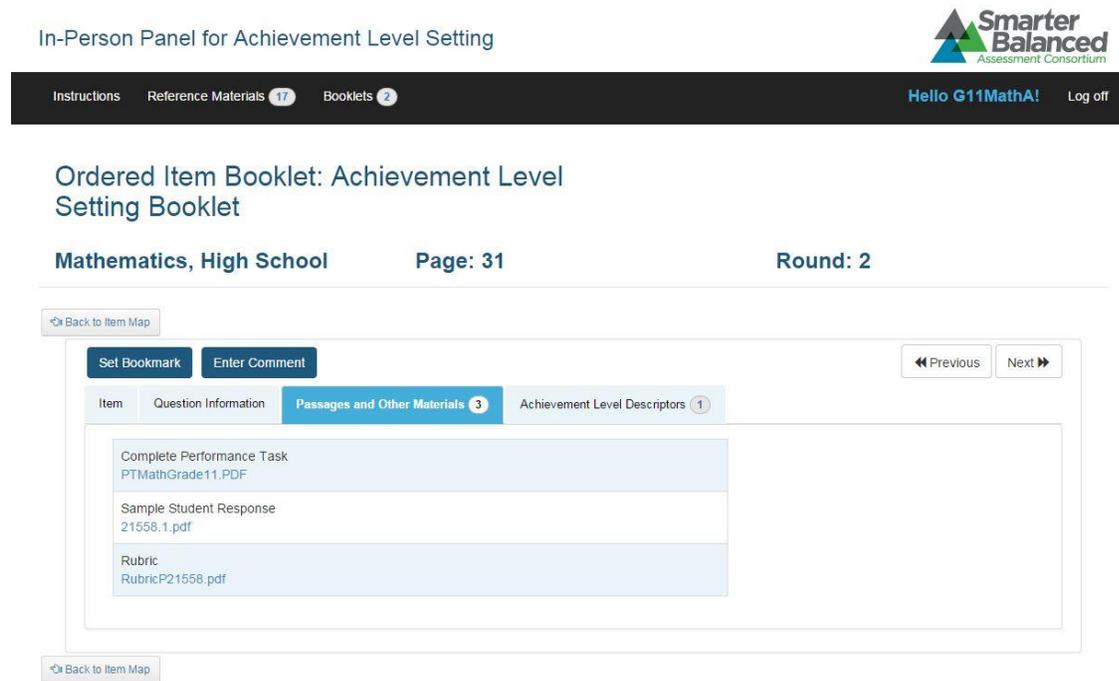
For Part 2, you are being asked to write a multi-paragraph article, so please be as thorough as possible. Type your response in the space provided. The box will expand as you type.

Remember to check your notes and your prewriting/planning as you write and then revise and edit your article.

**B I U T** [bulleted list] [numbered list] [indent] [outdent] [undo] [redo] [link] [unlink] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link] [insert image] [insert table] [insert video] [insert audio] [insert code] [insert link]

By clicking “Passages and Other Materials,” panelists could see resource materials such as reading or listening passages, sample student responses, and scoring rubrics, as shown in Figure 7.

Figure 7. OIB Page Showing Links to Performance Task, Sample Student Response, and Rubric.



In-Person Panel for Achievement Level Setting

Instructions Reference Materials 17 Booklets 2 Hello G11MathA! Log off

### Ordered Item Booklet: Achievement Level Setting Booklet

Mathematics, High School Page: 31 Round: 2

Back to Item Map

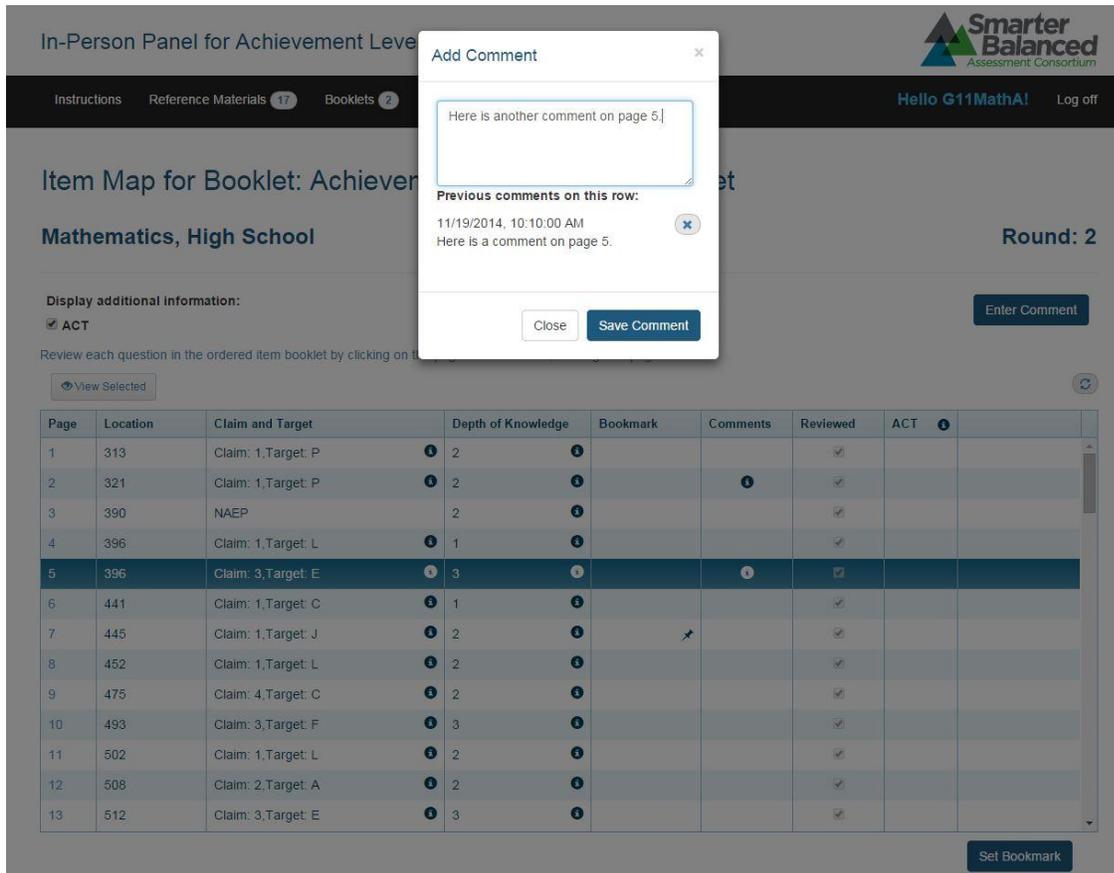
Set Bookmark Enter Comment Previous Next

Item	Question Information	Passages and Other Materials 3	Achievement Level Descriptors 1
Complete Performance Task		PTMathGrade11.PDF	
Sample Student Response		21558.1.pdf	
Rubric		RubricP21558.pdf	

Back to Item Map

The system was designed to allow panelists to leave comments on any test item by clicking on “Comments” in the OIB or in the appropriate row of the item map. These comments were intended to be used during inter-round discussions of the items by the in-person panelists or for the online panelists if they needed to leave the task and resume it later. Figure 8 illustrates the “Comment” function.

Figure 8. Comment.

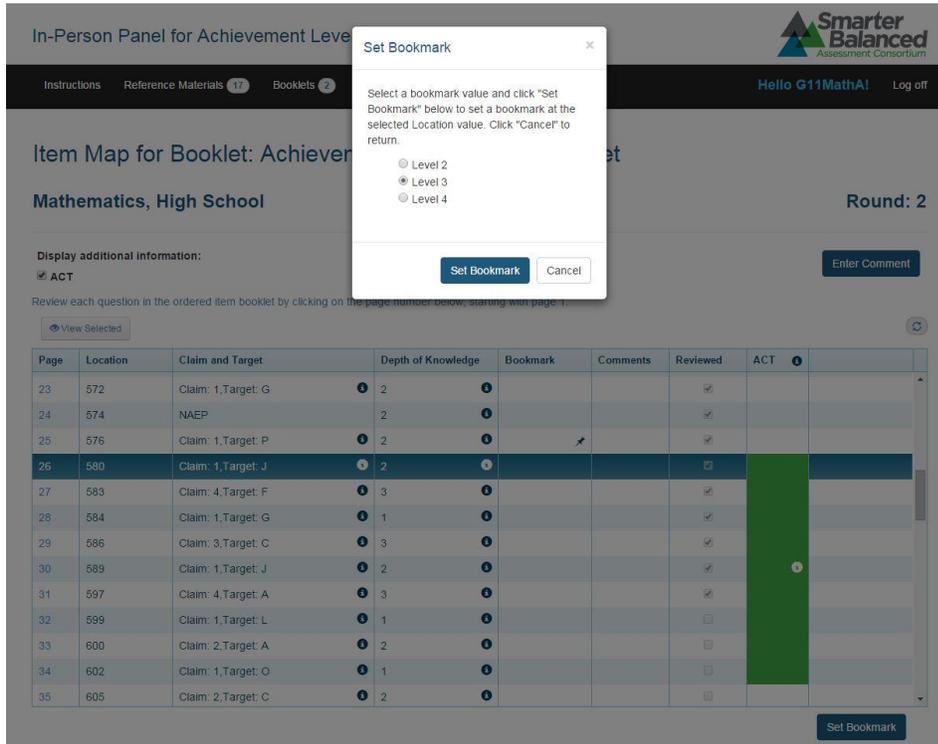


The screenshot shows the 'In-Person Panel for Achievement Level 3' interface. A modal dialog titled 'Add Comment' is open, allowing a user to enter a comment. The background shows the 'Item Map for Booklet: Achievement Mathematics, High School' with a table of items. The table has columns for Page, Location, Claim and Target, Depth of Knowledge, Bookmark, Comments, Reviewed, and ACT. Row 5 is highlighted, showing a comment icon in the 'Comments' column.

Page	Location	Claim and Target	Depth of Knowledge	Bookmark	Comments	Reviewed	ACT
1	313	Claim: 1, Target: P	2			<input checked="" type="checkbox"/>	
2	321	Claim: 1, Target: P	2			<input checked="" type="checkbox"/>	
3	390	NAEP	2			<input checked="" type="checkbox"/>	
4	396	Claim: 1, Target: L	1			<input checked="" type="checkbox"/>	
5	396	Claim: 3, Target: E	3			<input checked="" type="checkbox"/>	
6	441	Claim: 1, Target: C	1			<input checked="" type="checkbox"/>	
7	445	Claim: 1, Target: J	2			<input checked="" type="checkbox"/>	
8	452	Claim: 1, Target: L	2			<input checked="" type="checkbox"/>	
9	475	Claim: 4, Target: C	2			<input checked="" type="checkbox"/>	
10	493	Claim: 3, Target: F	3			<input checked="" type="checkbox"/>	
11	502	Claim: 1, Target: L	2			<input checked="" type="checkbox"/>	
12	508	Claim: 2, Target: A	2			<input checked="" type="checkbox"/>	
13	512	Claim: 3, Target: E	3			<input checked="" type="checkbox"/>	

After reviewing items, panelists could enter a bookmark by clicking either on the page in the OIB or in the appropriate row of the item map. Figure 9 illustrates the “Enter Bookmark” function. After entering all bookmarks (a single bookmark for Level 3 for the online panel activity or bookmarks for Levels 2, 3, and 4 for the in-person workshop), panelists were prompted to review their work and make sure they were ready to submit their bookmark(s), as shown in Figure 10.

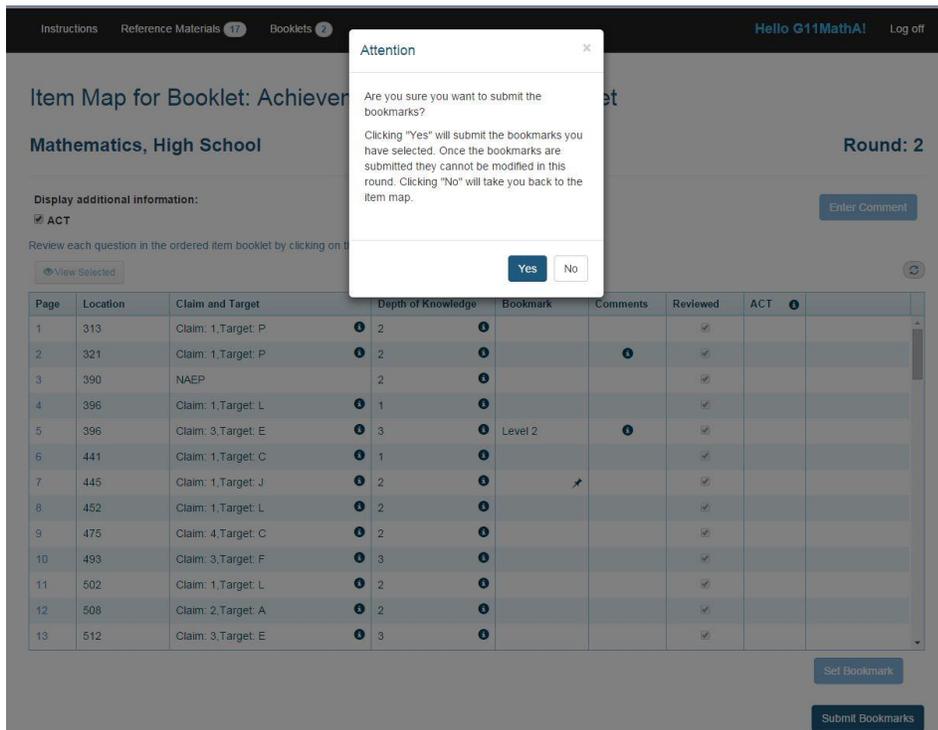
Figure 9. Set Bookmark Dropdown Box in the Item Map.



The screenshot shows the 'Set Bookmark' dialog box overlaid on the 'Item Map for Booklet: Achievement Mathematics, High School' interface. The dialog box contains the following text: 'Select a bookmark value and click "Set Bookmark" below to set a bookmark at the selected Location value. Click "Cancel" to return.' Below the text are three radio button options: 'Level 2', 'Level 3' (which is selected), and 'Level 4'. At the bottom of the dialog are two buttons: 'Set Bookmark' and 'Cancel'.

Page	Location	Claim and Target	Depth of Knowledge	Bookmark	Comments	Reviewed	ACT
23	572	Claim: 1, Target: G	2			<input checked="" type="checkbox"/>	
24	574	NAEP	2			<input checked="" type="checkbox"/>	
25	576	Claim: 1, Target: P	2			<input checked="" type="checkbox"/>	
26	580	Claim: 1, Target: J	2			<input checked="" type="checkbox"/>	
27	583	Claim: 4, Target: F	3			<input checked="" type="checkbox"/>	
28	584	Claim: 1, Target: G	1			<input checked="" type="checkbox"/>	
29	586	Claim: 3, Target: C	3			<input checked="" type="checkbox"/>	
30	589	Claim: 1, Target: J	2			<input checked="" type="checkbox"/>	
31	597	Claim: 4, Target: A	3			<input checked="" type="checkbox"/>	
32	599	Claim: 1, Target: L	1			<input type="checkbox"/>	
33	600	Claim: 2, Target: A	2			<input type="checkbox"/>	
34	602	Claim: 1, Target: O	1			<input type="checkbox"/>	
35	605	Claim: 2, Target: C	2			<input type="checkbox"/>	

Figure 10. Submitting Bookmarks.



The screenshot shows the 'Attention' dialog box overlaid on the 'Item Map for Booklet: Achievement Mathematics, High School' interface. The dialog box contains the following text: 'Are you sure you want to submit the bookmarks?' followed by a paragraph: 'Clicking "Yes" will submit the bookmarks you have selected. Once the bookmarks are submitted they cannot be modified in this round. Clicking "No" will take you back to the item map.' At the bottom of the dialog are two buttons: 'Yes' and 'No'.

Page	Location	Claim and Target	Depth of Knowledge	Bookmark	Comments	Reviewed	ACT
1	313	Claim: 1, Target: P	2			<input checked="" type="checkbox"/>	
2	321	Claim: 1, Target: P	2			<input checked="" type="checkbox"/>	
3	390	NAEP	2			<input checked="" type="checkbox"/>	
4	396	Claim: 1, Target: L	1			<input checked="" type="checkbox"/>	
5	396	Claim: 3, Target: E	3	Level 2		<input checked="" type="checkbox"/>	
6	441	Claim: 1, Target: C	1			<input checked="" type="checkbox"/>	
7	445	Claim: 1, Target: J	2			<input checked="" type="checkbox"/>	
8	452	Claim: 1, Target: L	2			<input checked="" type="checkbox"/>	
9	475	Claim: 4, Target: C	2			<input checked="" type="checkbox"/>	
10	493	Claim: 3, Target: F	3			<input checked="" type="checkbox"/>	
11	502	Claim: 1, Target: L	2			<input checked="" type="checkbox"/>	
12	508	Claim: 2, Target: A	2			<input checked="" type="checkbox"/>	
13	512	Claim: 3, Target: E	3			<input checked="" type="checkbox"/>	

## Design and Implementation of the Online Panel

The purpose of the online panel was to broaden the input into the process of making decisions about cut scores. In addition, the online panel allowed thousands of people to examine the tests and to express their opinions. The original proposal called for an online panel of 840 individuals; subsequent negotiations increased that number significantly. The final plan called for the contractor to support up to 250,000 online panelists. The intent was to have these individuals review a single ordered item booklet (OIB) and place a bookmark to indicate the location of the Level 3 cut score. The OIBs and support materials were the same as those used in the in-person workshop, but without the extensive training, interaction, and support provided to in-person workshop panelists. The online ALS differed from the typical Bookmark application described above in that panelists did not meet or receive successive rounds of feedback.

### *Online panel activities.*

Online panel activities commenced with recruitment (see Chapter 5), which began in April, 2014. Staff of McGraw-Hill Education (CTB), working in concert with Smarter Balanced staff and staff of Hager Sharp (H-S), crafted messages, first for educators and later for the general public, to alert them to the opportunity and explain the logistics.

Staff of Measurement Incorporated (MI) developed the software to support the online experience. That software included a home page, directions, links to reference materials, and digital OIBs, described below. Details of the software development and implementation are included in Appendix B.

Prior to the launch of the online panel on October 6, 2014, MI staff conducted a field test on August 14-15. That activity is described in Chapter 7 and summarized briefly here. Panelists for the online field test were 40 MI readers who logged in to a 30-minute webinar explaining the purpose of the activity and providing a brief introduction to the bookmark procedure. Panelists then had 48 hours to review an OIB for one of four ELA tests (grades 4, 6, 8, or 11) and enter a single bookmark, an activity that was estimated to take about three hours. Most who completed the activity took longer than three hours. Feedback from the panelists was collected via Survey Monkey, analyzed, and used to modify the process for October. Results are presented in Appendix D.

### *Conduct and results of the online panel.*

Online panelists signed up for one of six 48-hour windows, the first of which started on October 6. Ultimately, all windows were extended, and the final date was moved to October 18. By October 6, 10,099 individuals had registered to participate. Of that number, 5,840 logged in, and 2,660 placed a bookmark. **Online materials and software were deployed successfully, and capacity was more than adequate for application use.**

Results for online panelists entering a bookmark are presented in Table 2. Impact (percent of students who would score at or above the Level 3 cut score) is presented in Table 3. Impact is not reported for groups smaller than 25 online panelists. These results were also shared with the in-person workshop panelists and with the cross-grade review committee.

Table 2. Numbers of Online Panelists, by Role, Grade, and Subject

Grade	Teachers		Administrators		Higher Education		Other	
	ELA	Math	ELA	Math	ELA	Math	ELA	Math
3	151	167	67	37	9	5	31	30
4	89	124	31	28	2	4	16	22
5	96	114	31	35	5	5	12	21
6	66	91	11	22	4	8	9	17
7	70	100	12	22	4	5	6	8
8	87	115	27	39	4	7	11	22
11	193	267	55	64	60	83	13	26

Table 3. Impact of Online Panel Bookmark Placements: Percent of Students At or Above Level 3

Grade	Teachers		Administrators		Higher Education		Other	
	ELA	Math	ELA	Math	ELA	Math	ELA	Math
3	51%	54%	39%	50%			47%	45%
4	44%	43%	31%	52%				
5	61%	46%	65%	37%				
6	48%	38%						
7	57%	27%						
8	48%	18%	43%	18%				
11	55%	26%	48%	28%	56%	26%	58%	27%

The concept of an online panel is an innovation introduced to address the scale of the Smarter Balanced project and its number and variety of stakeholders. In addition to allowing wider

achievement level setting participation, the online panel approach promotes deeper understanding of the content standards and of the tasks used in schools. It also provided in-person panelists with feedback from a broader perspective. Online values for the level 2/3 cut were very similar to those for initial in-person values. This suggests that the approach should be explored in future standard setting venues in a manner that could provide wider participation and save on travel costs.

### **Design and Implementation of the In-Person Workshop**

As noted above, the bookmark procedure was used in the in-person workshop. The workshop took place at the Hilton Anatole in Dallas, Texas, on October 13-19, 2014. There were three waves of panels: the first wave, grade 11, began on Monday morning, October 13, and went through noon October 15; the second wave, grades 6–8, began on Wednesday morning, October 15, and went through noon October 17; the final wave, grades 3–5, began on Friday morning, October 17, and went through noon October 19. Table 4 summarizes the numbers of panelists by subject and grade. Table 5 summarizes the agenda for each 2.5-day session. Appendix D contains a detailed agenda for each day of the workshop.

Table 4. In-Person Workshop Panelists by Subject and Grade

Grade	English Language Arts/Literacy		Mathematics	
	Planned	Obtained	Planned	Obtained
3	1 panel of 30	1 panel of 26	1 panel of 30	1 panel of 30
4	1 panel of 30	1 panel of 27	1 panel of 30	1 panel of 29
5	1 panel of 30	1 panel of 27	1 panel of 30	1 panel of 29
6	1 panel of 30	1 panel of 30	1 panel of 30	1 panel of 30
7	1 panel of 30	1 panel of 27	1 panel of 30	1 panel of 30
8	1 panel of 30	1 panel of 30	1 panel of 30	1 panel of 29
11	2 panels of 36	2 panels of 34	2 panels of 36	2 panels of 35
Total	252	235	252	247
Grand Total	504	482 (95.6%)		

Table 5. High-Level Agenda for Each In-Person Workshop.

Day - Time	Event(s)
Day 1 A.M.	Welcome; overview, training on CCSS, ALDs, tests
Day 1 P.M.	Review of Ordered Item Booklet
Day 2 A.M.	Orientation to the Bookmark Procedure; complete Round 1
Day 2 P.M.	Review Round 1; complete Round 2
Day 3 A.M.	Review Round 2; complete Round 3; evaluate process

#### ***Recruitment and selection of panelists.***

Recruitment of panelists for the In-Person Workshop began April 15. K-12 State Leads, Higher Education Leads, and Teacher Involvement Coordinators received communication tools developed by the contractor and approved by Smarter Balanced to enable them to recruit teachers (general as well as teachers of English language learners and students with disabilities), school administrators, higher education faculty, business and community leaders, and parents. Each Smarter Balanced state had 20–25 positions to fill, giving each state an opportunity to have at least one representative for each of the 14 tests.

#### ***Preparation of materials.***

Staff of MI and CTB prepared the following training materials, all of which can be found in Appendix C:

- Introductory PowerPoint presentation to orient panelists to the goals and tasks of the workshop
- Common Core State Standards – up-to-date versions of the subject/grade-specific content standards as well as guidelines to their use in the achievement level setting activity
- Achievement Level Descriptors – up-to-date versions of the ALDs for the specific subject and grade for each panel
- Practice Test – using the online version of the Smarter Balanced practice tests for each grade and subject
- Orientation to the ordered item booklet – PowerPoint presentation designed to show panelists what to look for and questions to ask as they reviewed items in the OIB
- Orientation to the Bookmark procedure – PowerPoint presentation designed to show panelists how Bookmark works and specifically how panelists were to implement the procedure in a computer-based environment
- Bookmark Orientation Round – an exercise involving a 6-page OIB that panelists reviewed prior to entering a single bookmark and discussing their placements in a large-group setting.
- Readiness Form – a multipart form that asked panelists at several key points during the process how well they understood the process they were implementing and how ready they were to proceed to the next step
- Evaluation Form – a series of statements about the training, environment, and conduct of the workshop that the panelists responded to on a graded scale (such as Strongly Agree to Strongly Disagree)

MI and CTB staff drafted all training materials and submitted them to Smarter Balanced staff and the external auditor for review in advance of the workshop. Final versions of all training materials reflect the comments and recommendations of these reviews and were approved by Smarter Balanced leadership prior to use. All training materials are included in Appendix C.

#### ***Training of facilitators and table leaders.***

In advance of the in-person workshop, staff of MI and CTB prepared a detailed facilitator script which was reviewed and approved by Smarter Balanced. Staff identified as facilitators studied the scripts and participated in in-house training sessions the week prior to the in-person workshop. In addition, Mr. Ricardo Mercado of CTB conducted a two-hour facilitator training session on Sunday night, October 12, on Tuesday night, October 14, and on Thursday night, October 16, as facilitators for each wave arrived in Dallas. At the same time, Dr. Jennifer Lord-Bessen of CTB provided a two-hour orientation for table leaders who had been identified in advance by their State Leads. Training materials for those sessions are included in Appendix C.

#### ***Orientation and training.***

Using the training materials approved by Smarter Balanced, MI and CTB staff provided large-group and small-group training. For the opening session, Dr. Joe Willhoft gave the welcome and charge. Dr. Michael Bunch of MI provided specific training on the content standards, ALDs, and practice tests. Dr. Daniel Lewis of CTB provided the orientation to the Bookmark procedure. At the end of each training session, panelists completed a portion of the Readiness Form (see Appendix C).

In-Person Workshop panelists were encouraged to review the appropriate ALDs and CCSS standards prior to coming to the workshop. However, it was not assumed that all had done so, and panelists

were given an opportunity not only to review the materials on site but to discuss them in a large-group setting. They had an opportunity to indicate on the Readiness Form just how familiar they were with those materials. No panelist was permitted to advance to item review without indicating familiarity with the ALDs and content standards and indicating readiness to proceed.

The afternoon of Day 1 was devoted entirely to review of the OIB. In addition to being oriented to the software, panelists were introduced to the test items themselves. They spent the entire afternoon annotating items, using the Comments function of the software, and discussing items with others at their table in terms of the first two guiding questions. While this activity had been scheduled to end at 5 p.m. on Day 1, all panels required additional time and received from 30 to 60 minutes to complete the task at the beginning of Day 2, following orientation to the bookmark procedure.

At the beginning of Day 2, all panelists assembled in the ballroom for orientation to the Bookmark procedure. Dr. Daniel Lewis, Chief Research Advisor at CTB and co-creator of the Bookmark procedure, provided the orientation and answered questions. Following the orientation to the Bookmark procedure, panelists adjourned to their small groups to gain first-hand experience in setting a bookmark through a practice exercise. This exercise consisted of a 6-page OIB with items of varying difficulty. Each panel had access to two facilitators who oriented panelists to the computers and software and showed them how to navigate the OIB. Panelists then had several minutes to review the six items and enter a bookmark. The facilitator then led a discussion focusing on how many panelists chose each page to place their bookmarks. Following this discussion, panelists completed a section of their Readiness Forms, indicating their readiness to begin Round 1.

#### *Round-by-round item review and discussion.*

Panelists were invited to work through their on-screen OIBs and discuss the items with others at their table. They were able to discuss their opinions with one another at their table as much as they wished, but when they entered a bookmark, it was to be their bookmark, not that of the table. They started by placing a bookmark for Level 3, then Level 4, and finally, Level 2. After placing three bookmarks, panelists were dismissed for lunch, during which time CTB staff tallied bookmarks but did not provide reports to the panelists. Results are shown in Table 6 in terms of median bookmark placement for each subject, grade, and level. Complete results, including distributions of bookmark placements, are included in Appendix D.

Table 6. Results of Round 1 of Bookmark Placement (Entries are Median Page Numbers).

Subject/Grade	ELA			Math		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
ELA 3	16.0	38.0	58.5	22.0	47.0	69.5
ELA 4	20.0	42.0	60.0	12.0	33.0	69.0
ELA 5	13.0	27.0	63.0	21.5	50.0	65.5
ELA 6	15.0	35.0	63.0	18.0	37.5	61.0
ELA 7	16.0	41.0	69.0	21.5	42.5	63.0
ELA 8	19.0	39.5	68.0	18.0	39.0	58.0
EALA 11	21.5	45.0	66.0	19.0	48.5	69.0

Panelists, upon returning from lunch, were directed to share their Round 1 bookmark placements with others at their table, discuss their rationales for placing those bookmarks, and compare

approaches as well as comments they had left on the item map. The facilitator then introduced and led a discussion on the bookmark placements of the online panel. Once they completed their discussions, panelists completed the portion of the Readiness Form that indicated they were ready to begin Round 2.

In Round 2, panelists proceeded as in Round 1, conferring with others at their table but entering their own bookmarks. When they entered three bookmarks and submitted them, they were free to log out for the day. Results of Round 2 are shown in Table 7. Complete results, including bookmark distributions and interquartile ranges, are shown in Appendix D.

Table 7. Results of Round 2 of Bookmark Placement (Entries are Median Page Numbers).

Grade	ELA			Math		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
3	19.0	38.0	57.5	28.0	49.0	70.0
4	20.0	44.0	63.0	9.0	32.0	71.0
5	14.0	27.0	61.0	20.0	50.5	64.0
6	15.0	36.5	63.0	16.5	37.0	60.0
7	16.0	37.5	69.5	18.0	46.0	61.0
8	17.0	40.5	66.5	17.0	40.0	60.0
11	22.0	42.0	65.0	20.0	50.0	69.0

Panelists returned the morning of the third day to see the results of Round 2. The facilitator led a discussion of the range of bookmark placements, corresponding cut scores, and percentages of students classified at each level, based on the Round 2 cut scores. Once again, the facilitator showed the online panel results, this time in terms of percentages of students at or above Level 3, based on online panelists’ bookmark placements. A room-wide discussion ensued. Finally, facilitators revealed the impact for the next grade up; i.e., panelists in grade 8 were able to see the final impact of the cut scores set by grade 11 panels, panelists in grade 7 were able to see the Round 2 results for grade 8, and so on down to grade 3. By virtue of being first, grade 11 panelists did not get to see results of any other in-person workshop panels.

After review and discussion of all results, panelists completed the final section of their Readiness Forms and began Round 3. They completed Round 3 as they had Round 2, bypassing many pages which no one had recommended in previous rounds and keeping or changing their bookmark placements depending on their response to the discussion. Each panelist entered three bookmarks and then submitted those bookmarks for analysis. Results of Round 3 are shown in Tables 8 (bookmark placement) and 9 (scale score cuts and percentages of students at or above each level).

Table 8. Results of Round 3 of Bookmark Placement (Entries are Median Page Numbers).

Subject/Grade	ELA			Math		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
3	13.0	33.0	54.0	27.0	47.0	70.0
4	19.0	43.0	62.0	15.0	39.0	71.0
5	11.0	27.0	63.0	19.0	50.0	64.0
6	14.5	34.5	60.5	18.0	45.5	61.5
7	16.0	38.0	66.0	17.0	45.0	61.0
8	18.0	39.5	68.0	16.0	40.0	60.0
11	19.0	42.0	65.0	19.5	48.0	68.0

Table 9. Round 3 Cut Score Recommendations: Scale Score Cuts and % At or Above

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 3 English Language Arts/Literacy	362	66.5%	427	40.1%	485	19.1%
Grade 3 Mathematics	383	67.3%	436	38.9%	506	10.8%
Grade 4 English Language Arts/Literacy	413	64.4%	470	42.0%	530	18.9%
Grade 4 Mathematics	400	77.6%	470	44.7%	541	15.6%
Grade 5 English Language Arts/Literacy	406	78.7%	450	64.0%	574	16.9%
Grade 5 Mathematics	459	63.5%	532	31.4%	583	13.8%
Grade 6 English Language Arts/Literacy	466	66.6%	527	42.2%	614	12.2%
Grade 6 Mathematics	491	58.3%	561	29.4%	603	15.6%
Grade 7 English Language Arts/Literacy	474	68.2%	547	40.1%	660	6.6%
Grade 7 Mathematics	513	53.1%	609	19.3%	674	5.8%
Grade 8 English Language Arts/Literacy	471	76.4%	543	50.9%	663	10.2%
Grade 8 Mathematics	534	51.3%	605	25.6%	683	7.4%

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 11 English Language Arts/Literacy	490	72.9%	565	47.6%	677	12.1%
Grade 11 Mathematics	533	62.6%	644	28.0%	740	8.0%

After entering their Round 3 bookmarks, panelists took a short break and returned to review the final cut scores, complete a final questionnaire, and then evaluate the process, using online evaluation forms. Results of those questionnaires and evaluation forms are summarized in Tables 10 and 11. Questionnaire and evaluation results for individual panels are included in Appendix D.

Table 10. Round 3 Questionnaire Results: Confidence in Cut Scores Recommended (Discounting Blanks)

*How confident are you about the three bookmarks you just entered?*

Bookmark	Very Confident	Confident	Uncertain	Very Uncertain	Total
Level 2	222 (47%)	237 (51%)	10 (2%)	0 (0%)	469
Level 3	234 (50%)	220 (47%)	15 (3%)	0 (0%)	469
Level 4	245 (52%)	217 (46%)	7 (1%)	0 (0%)	469

Table 11. Summary of Round 3 Evaluation Responses (Discounting Blanks)

Evaluation Statement	Strongly Agree	Agree	Disagree	Strongly Disagree	Total
The orientation provided me with a clear understanding of the purpose of the meeting.	253 (58%)	170 (39%)	13 (3%)	2 (0%)	438
The workshop leaders clearly explained the task.	245 (56%)	161 (37%)	25 (6%)	7 (2%)	438
The training and practice exercises helped me understand how to perform the task.	247 (56%)	174 (40%)	16 (4%)	1 (0%)	438
Taking the practice test helped me to understand the assessment.	231 (53%)	192 (44%)	14 (3%)	1 (0%)	438
The Achievement Level Descriptions were clear and useful.	199 (45%)	216 (49%)	21 (5%)	2 (0%)	438
The large and small group discussions aided my understanding of the process.	300 (68%)	132 (30%)	4 (1%)	2 (0%)	438
The time provided for discussions was appropriate.	230 (53%)	184 (42%)	23 (5%)	1 (0%)	438

Evaluation Statement	Strongly Agree	Agree	Dis-agree	Strongly Disagree	Total
There was an equal opportunity for everyone in my group to contribute his/her ideas and opinions.	292 (67%)	135 (31%)	8 (2%)	3 (1%)	438
I was able to follow the instructions and complete the rating tasks accurately.	284 (65%)	151 (34%)	1 (0%)	2 (0%)	438
The discussions after the first round of ratings were helpful to me.	273 (62%)	151 (34%)	12 (3%)	2 (0%)	438
The discussions after the second round of ratings were helpful to me	270 (62%)	156 (36%)	11 (3%)	1 (0%)	438
The information showing the distribution of student scores was helpful to me.	220 (50%)	200 (46%)	13 (3%)	4 (1%)	437
I am confident about the defensibility and appropriateness of the final recommended cut scores.	203 (46%)	202 (46%)	27 (6%)	6 (1%)	438
The facilities and food service helped create a productive and efficient working environment.	324 (74%)	104 (24%)	10 (2%)	0 (0%)	438

### *Data analysis and reporting.*

As panelists entered and submitted bookmarks, the data flowed directly from their computers to servers MI had set up prior to the start of the workshop. Staff from CTB, using BookmarkPro software, received the data, analyzed them, and produced reports that facilitators shared at the beginning of the next round. A full set of reports is included in Appendix D.

### *Design and Implementation of the Cross-Grade Review Committee*

The vertical articulation committee was renamed the cross-grade review committee to reflect more clearly the nature of their task, which was to review all cut scores and impact across all grades within a given subject and make adjustments where necessary to prevent or minimize large discontinuities in impact across grades. For example, if 50 percent of students in grades 5, 6, and 8 were at or above Level 3, but only 40 percent of grade 7 students were above Level 3, such a discrepancy would need to be examined.

The committees (32 members each for ELA and mathematics) met on October 20, 2014. Dr. Bunch provided an introduction to the tasks and ground rules. The complete presentation is included in Appendix B and is summarized here.

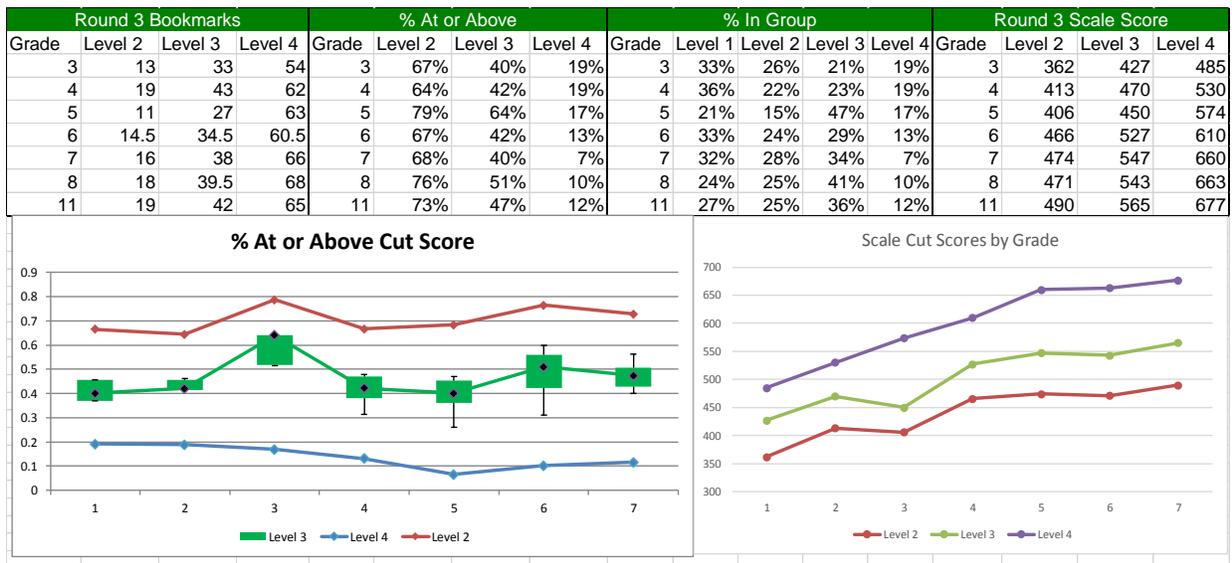
Trends in percentages of students at or above a given level tend to be of one of three types: either more and more students reach a given level over time or across grades (i.e., generally increasing), the same percentages of students reach a given level year after year or from one grade to the next (level), or fewer and fewer students reach a given level over time or across years (generally decreasing). Trends that go up one year and down the next, or up for one grade and down for the

next, are much more difficult to explain (though there may be legitimate reasons for such trends). The task of the cross-grade review committee was to investigate any discontinuities and determine whether they were accurate reflections of reality or indications that one or more panels had been overly stringent or overly lenient.

Dr. Bunch explained that the process would include review of actual OIBs and impact data, starting with grades 8 and 11. Any panelist would be welcome to recommend changing any cut score, although a panelist from the grade directly involved or from an adjacent grade would be the preferred initiator of any recommended change. He explained the process for introducing and seconding a motion to change a cut score, to be followed by discussion and a vote. Given that any change would alter the work of a panel of 30 to 36 people, a 2/3 super majority was required to pass any recommended change.

After the orientation, 32 mathematics panelists reconvened in an adjacent room, while the 32 ELA panelists remained in the room in which the orientation had taken place. In both rooms, computers from the previous week’s in-person workshop were still in place with all software still loaded. For each subject, all seven OIBs and all support materials used by in-person workshop panelists were available. Whenever anyone suggested a change, the facilitator (Dr. Bunch for ELA and Dr. Lewis for mathematics) was able to show on a large screen in the front of the room a projected image of how that change would affect impact. An example of the on-screen graphic is shown in Figure 11.

Figure 11. Cross-Grade Review Graphic.



In Figure 11, the four tables at the top represent the Round 3 median bookmark placements, the percentages at or above Levels 2-4 based on those bookmark placements, the resulting percentages of students classified into each level, and the Round 3 bookmark placements translated into temporary scale scores. The graph on the bottom left reflects the impact in the second table, with the black dots representing the medians, the green boxes representing the interquartile ranges, and the black vertical lines representing the 10<sup>th</sup> and 90<sup>th</sup> percentiles. Panelists could recommend changing any bookmark placement in the first table, and all other tables, as well as the two graphs at the bottom, would immediately change accordingly.

Panelists began by reviewing cut scores for grades 8 and 11 and then worked their way down through the middle and elementary grades. By the end of the day, the ELA committee had made 8 changes, and the mathematics committee had made 11. Final results for the two committees are shown in Table 12, with changes from Round 3 highlighted in yellow.

Table 12. Cross-Grade Review Results

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 3 English Language Arts/Literacy	362	66.5%	427	40.1%	485	19.1%
Grade 3 Mathematics	381	68.3%	436	38.9%	501	12.1%
Grade 4 English Language Arts/Literacy	413	64.4%	470	42.0%	530	18.9%
Grade 4 Mathematics	413	72.3%	487	36.5%	551	12.6%
Grade 5 English Language Arts/Literacy	434	69.7%	494	47.1%	574	16.9%
Grade 5 Mathematics	459	63.5%	532	31.4%	583	13.8%
Grade 6 English Language Arts/Literacy	453	71.3%	527	42.2%	614	12.2%
Grade 6 Mathematics	491	58.3%	570	26.1%	609	14.0%
Grade 7 English Language Arts/Literacy	474	68.2%	547	40.1%	644	9.5%
Grade 7 Mathematics	513	53.1%	596	23.2%	674	5.8%
Grade 8 English Language Arts/Literacy	482	73.1%	562	43.3%	663	10.2%
Grade 8 Mathematics	534	51.3%	616	22.1%	683	7.4%
Grade 11 English Language Arts/Literacy	488	73.3%	578	42.8%	677	11.6%
Grade 11 Mathematics	565	48.3%	650	26.4%	740	5.8%

### Approval by Chiefs

Subsequent to the completion of the cross-grade review, Smarter Balanced and contractor staff prepared to present results to the Chiefs for review and approval. On November 6, Chiefs met in Chicago to review the results. While endorsing the work of the panels, the Chiefs did not vote on the cut scores. A second meeting was scheduled for November 14, in conjunction with the meeting of the Council of Chief State School Officers (CCSSO) in San Diego. Meanwhile, Smarter Balanced staff prepared options to present to the Chiefs at the November 14 meeting, incorporating evidence from recent studies conducted by the National Assessment Governing Board (NAGB). In addition, Smarter Balanced staff created a new reporting scale, replacing the temporary scale used throughout achievement level setting and cross-grade review. While the temporary scale had a range of 100 to 900, the final scale had a range of 2000 to 3000 and can be easily derived from the temporary scale by adding 2000 to the original scale. Thus, for example, the grade 11 mathematics Level 2 cut score of 565 would translate to a final score of 2565.

The two options presented to the Chiefs at the November 14 meeting consisted of the results shown in Table 12 and those same results moderated in the direction of the NAGB results. Specifically, while working within a range of plus-or-minus one standard error of measurement of the cut scores recommended by the cross-grade review committee, Smarter Balanced staff recommended ELA cut scores that were higher and mathematics cut scores that were lower than those recommended by the cross-grade review committee. These modifications kept recommended cut scores within or very close to the one SEM range, approximated NAGB results, and brought ELA and mathematics impacts into closer alignment with each other. The Chiefs voted unanimously (with two abstentions) on November 14 to approve the modified cut scores, presented in Table 13.

Table 13. Final Cut Scores Approved By Chiefs, With Impact Data.

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 3 English Language Arts/Literacy	2367	65%	2432	38%	2490	18%
Grade 3 Mathematics	2381	68%	2436	39%	2501	12%
Grade 4 English Language Arts/Literacy	2416	63%	2473	41%	2533	18%
Grade 4 Mathematics	2411	73%	2485	37%	2549	13%
Grade 5 English Language Arts/Literacy	2442	67%	2502	44%	2582	15%
Grade 5 Mathematics	2455	65%	2528	33%	2579	15%
Grade 6 English Language Arts/Literacy	2457	70%	2531	41%	2618	11%
Grade 6 Mathematics	2473	65%	2552	33%	2610	14%
Grade 7 English Language Arts/Literacy	2479	66%	2552	38%	2649	8%
Grade 7 Mathematics	2484	64%	2567	33%	2635	13%
Grade 8 English Language Arts/Literacy	2487	72%	2567	41%	2668	9%
Grade 8 Mathematics	2504	62%	2586	32%	2653	13%
Grade 11 English Language Arts/Literacy	2493	72%	2583	41%	2682	11%
Grade 11 Mathematics	2543	60%	2628	33%	2718	11%

## Achievement Level Descriptors

Prior to the awarding of a contract for achievement level setting, Smarter Balanced had awarded several other contracts for program management, test development, and development of achievement level descriptors (ALDs). There are, or will be by spring 2015, four sets of ALDs (from Egan, Schneider, & Ferrara, 2012):

- Policy – brief statements that articulate policy makers’ vision of goals and rigor for the final performance standards;
- Range – guidelines created by test developers to identify which aspects of items align to a particular performance level with regard to the cognitive and content rigor that has been defined;
- Threshold (Target) – detailed statements created in conjunction with the Range ALDs and used by achievement level setting panelists to represent the knowledge and skills of a student just at the threshold of a given level;
- Reporting – relatively brief statements developed by a sponsoring agency once cut scores are finalized, to define the appropriate and intended interpretations of test scores.

Policy ALDs allowed Smarter Balanced to communicate to the educational community its intentions for development and implementation of rigorous assessments. The Range ALDs were used to guide item developers. Threshold ALDs were used to guide online and in-person achievement level setting panelists in the placement of bookmarks to recommend cut scores. In the spring of 2015, Smarter Balanced will use the reporting ALDs to describe the achievement of millions of students to parents, schools, districts, and states.

Once final cut scores are set, it is advisable to review ALDs to make sure that they are aligned to the cut scores. This section documents revisions to the Threshold ALDs in light of modifications to cut scores recommended by panelists, subsequent review of the Range ALDs, and development of the Reporting ALDs.

### Threshold ALDs

Threshold ALDs were a central part of the training of online and in-person achievement level setting panelists. Through three rounds of achievement level setting, which included review of recommendations of Online Panelists, In-Person Workshop panelists justified each cut score on the basis of the content alignment of a test item on a given page of an ordered item booklet (OIB) with the description in the Threshold ALD. Subsequent changes to those cut scores by the Cross-Grade Review Committees (formerly known as the Vertical Articulation Committees) were also grounded in the Threshold ALDs. All cut score recommendations going forward to the Chiefs were thus firmly grounded in the language of the Threshold ALDs.

The final cut scores, however, were not always the same as those emerging from the Cross-Grade Review Committees. In some instances, they went up; in others, they went down. Thus, a review of the alignment of the final cut scores and the Threshold ALDs was in order. The process, findings, and recommendations are detailed below.

**Comparison of final cuts to recommended cuts.**

Table 14 compares final cut scores to those recommended by the Cross-Grade Review Committees. In each instance, cut scores have been translated into page numbers in the OIBs, since these had been the focus of the initial recommendations. The plausible range of page numbers indicates the interquartile range of bookmark placements from Round 3 of the In-Person Workshop, as augmented by the Cross-Grade Review Committee. For example, if the middle 50 percent of the range of bookmarks for Level 3 placed by panelists in the In-Person Workshop for grade 4 Mathematics was 38 to 48, but the Cross-Grade Review Committee moved the bookmark to page 49, the plausible range was from pages 38 to 49. If the Cross-Grade Review Committee did not alter the cut score or moved it within the Round 3 range, the plausible range was whatever it had been at the end of Round 3. In 14, E and M refer to English language arts/literacy and Mathematics, and L2 – L4 refer to Levels 2 – 4. Cell entries are OIB page numbers. For the Plausible Range, medians and interquartile ranges can include page numbers that are not whole numbers; thus, they are reported in quarter-page increments.

Table 14. Comparison of Final Cuts to Those Recommended by the Cross-Grade Review Committees.

		OIB Page #			Plausible Range						In Range?			Out of Range by ___ Pages		
		L2	L3	L4	L2		L3		L4		L2	L3	L4	L2	L3	L4
Subject	Grade	L2	L3	L4	From	To	From	To	From	To						
E	3	14	36	57	11.75	15.50	28.00	38.00	53.00	61.75	Yes	Yes	Yes			
E	4	21	44	63	15.00	20.00	37.00	44.00	60.00	63.00	No	Yes	Yes	1		
E	5	21	38	65	10.00	18.00	27.00	37.00	61.00	65.00	No	No	Yes	3	1	
E	6	11	35	61	7.75	19.00	29.00	40.00	52.00	63.25	Yes	Yes	Yes			
E	7	16	39	65	8.00	16.00	34.50	43.50	64.00	74.00	Yes	Yes	Yes			
E	8	22	46	68	14.00	21.50	34.00	46.50	60.00	70.00	No	Yes	Yes	0.5		
E	11	19	46	65	15.25	23.00	40.00	45.00	63.00	66.00	Yes	No	Yes		1	
M	3	26	46	69	26.00	28.00	44.50	53.00	66.00	72.25	Yes	Yes	Yes			
M	4	17	46	72	12.00	18.00	38.00	49.00	71.00	73.00	Yes	Yes	Yes			
M	5	17	49	61	18.25	21.00	50.00	51.00	62.00	64.00	No	No	No	-1.25	-1	-1
M	6	15	40	63	13.00	20.00	32.50	53.50	59.00	63.00	Yes	Yes	Yes			
M	7	9	30	53	13.25	21.00	40.00	51.00	58.75	64.00	No	No	No	-4.25	-10	-5.75
M	8	8	33	50	15.00	18.00	36.50	48.00	57.50	63.00	No	No	No	-7	-3.5	-7.5
M	11	20	44	63	17.50	27.00	44.00	55.25	63.75	69.00	Yes	Yes	No			-0.75

As can be seen, most final cuts are either in range or very close to the plausible range. For those not within the Plausible Range, the final three columns indicate the distance from the edge of the range. Those out-of-range cuts were the primary focus of the ALD review.

**ALD review.**

MI staff drafted a plan, based on Table 14, and presented it to Smarter Balanced staff on November 24. Smarter Balanced approved the plan, and MI set it into motion. MI staff reviewed threshold ALDs, test blueprints, panelist and facilitator notes from the In-Person Workshop Panel and Cross-Grade Review Committees, and comments from the Online Panel for all cut scores in Table 14 that were out of range. They then shared their finding with Smarter Balanced staff, who reviewed them and provided feedback. MI staff then submitted final recommendations to Smarter Balanced staff for review and approval. Those findings and recommendations are detailed in the next subsection.

### **Findings and recommendations.**

MI content specialists reviewed five modified cut scores for English language arts/literacy and ten for mathematics. In each instance, the content specialists were able to justify the new cut score in terms of the threshold ALDs. In several instances, the new cut score was only a page or two away from the plausible range established by the cross-grade review committee. However, even when the new cut score was as much as 10 pages below the range, the content specialists found that the content of the item associated with the new cut score met the threshold ALD criteria; i.e., that the item just below the bookmark presented a student at the threshold with about a 50% chance of answering correctly and that the item at the bookmark presented the student at the threshold with less than a 50% chance of answering correctly. Thus, in 15 out of 15 instances, the content specialists concluded that the final cut scores aligned to the threshold ALDs. An item-by-item account of the findings and recommendations is included in a separate report.

### **Range ALDs**

As noted above, the purpose of Range ALDs is to guide test item developers. Specifically, item developers need to know what is to be expected of students at Levels 1, 2, 3, and 4. If those expectations change, item-development guidance also needs to change. As item development will continue into the foreseeable future, any change in expectations of students at various levels, as reflected in the Threshold ALDs, needs to be reflected in the Range ALDs. However, given that there were no changes to the Threshold ALDs, no changes are recommended for the Range ALDs.

### **Reporting ALDs**

As noted above, reporting ALDs are relatively brief statements developed by a sponsoring agency once cut scores are finalized, to define the appropriate and intended interpretations of test scores. Ideally, they should reflect the specific knowledge, skills, and processes embodied in the tests. However, in the case of computer adaptive tests, those sets may vary from student to student; therefore, the reporting ALDs for Smarter Balanced will need to be more generic, reflecting a range of knowledge, skills, and processes.

MI staff gathered requirements and recommendations from Smarter Balanced staff and others and drafted a matrix of policy ALDs and reporting ALDs for high school, grades 6-8, and grades 3-5. Staff of MI, CTB, and Hager Sharp met on December 1 to review the matrix and make revisions. This revised matrix was presented to Smarter Balanced leadership on December 2 for further review and revision. The results of that presentation were forwarded to Smarter Balanced for further revision. Draft reporting ALDs are included in a separate report.

### **Long Range Validity Agenda for Performance Level Cut Scores**

As Smarter Balanced shifts from a developmental to an operational mode, additional research on cut scores is planned. The final task under Contract 21 is to prepare a long-range research agenda that will allow Smarter Balanced to test the validity of the cut scores against various external criteria.

In 2012, Smarter Balanced commissioned Stephen Sireci to prepare a comprehensive research “to demonstrate that the assessment system adheres to professional and Federal guidelines for fair and high quality assessment...to provide a comprehensive and detailed research agenda for the Consortium that includes suggestions and guidance for both short- and long-term research activities that will support Consortium goals” (Sireci, 2012, p. 5). The current report has a much more narrow focus: validation of cut scores established in the fall of 2014. However, the Sireci (2012) research agenda provides a solid foundation on which to build the plan for cut score validation.

The present plan is further guided by the 2014 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), just as the Sireci agenda was guided by the 1999 *Standards*, which are similar in many ways to the 2014 *Standards*. The present proposal is also guided by *Peer Review Guidance* (U. S. Department of Education, 2009), principles of Universal Design (Johnson, Altman, & Thurlow, 2006), Michael Kane’s recent essays on validation (Kane, 2001, 2006), and similar work by Susan Loomis (2011). In particular, this paper (as does the Sireci paper) uses the theoretical framework and terminology employed by Kane (2001, 2006) and reflected in the 2014 *Standards*; i.e., “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA, APA, NCME, 2014, p. 11). With specific reference to achievement level setting, Kane (2001, p. 54) notes, “To set a standard is to develop policy, and policy decisions are not right or wrong. They can be wise or unwise, effective or ineffective, but they cannot be validated by comparing them to some external criterion. The argument for validity, or appropriateness, of a standard is necessarily extended, complex, and circumstantial.”

The validation studies described in this research plan focus principally on the summative assessments (i.e., those for which cut scores are to be set) and represent the perspective of the contractor for Smarter Balanced Contract 21 (Achievement Level Setting). They are based, in large measure, on previous large-scale cut-score validation efforts that involved collection and analysis of a wide range of data external to the assessment programs in question; i.e., the American Diploma Project (Miles, Beimers, & Way, 2010).

There will be opportunities following the 2015 operational test administration and beyond to examine the cut scores with respect to **internal** variables from the Smarter Balanced assessment system and targeted **external** variables. Appraising the cut scores from these various perspectives will yield important information as to the appropriateness of the interpretations and uses of these scores. The remainder of this section will therefore be devoted to a description of validation studies with internal variables and validation studies with external variables.

### Validation Studies with Internal Variables

Internal variables are those specific to the tests themselves, their blueprints, their internal structures, and their score distributions. Specific variables and uses are outlined below.

- *Interim assessments*: States and districts will have administered the interim assessments and collected test score data based on the performance of their students. Because these assessments are on the same scale as the summative test, it will be possible to calculate the agreement level of students’ achievement levels that emerge from the interim and summative scores.
- *Formative assessments*: Using samples of students matched on propensity scores, it will be possible to examine the performance of students on the interim and summative assessments based on their use of the formative assessments.

- *Digital Library:* Using samples of students matched on propensity scores, it will be possible to examine the performance of students on the interim and summative assessments based on their use of the Digital Library materials.
- *Expert studies:* Blueprints for the 2015 assessments were approved at the Collaborative Conference on April 30. Given the computer-adaptive nature of the 2015 assessments, there will not be a static form whose alignment to a blueprint can be readily evaluated. However, it will be possible to assemble proto-tests (i.e., collections of items for students at specific hypothetical ability levels) that will closely resemble tests administered to actual students. Those proto-tests should be reviewed by content experts. Specifically, higher education faculty should evaluate the high school tests, high school faculty will evaluate middle school tests, and middle school faculty should evaluate elementary school tests for alignment to the blueprints as well as depth of knowledge and rigor.
- *Level on subsequent tests:* Just as one would expect some degree of consistency of impact across grades, one would also expect individual students to perform consistently from one grade to the next. Specifically, most students would be expected to progress from their initial level to higher levels over time. Failure of large numbers of students to progress as expected would call into question the appropriateness of the cut scores, whether instruction and motivation were adequate, or whether some combination of factors caused the deviation from expectation. A longitudinal study in which a sample of students in grades 3-8, over a three-year period (until a majority of the first year's eighth graders take the high school tests) is recommended. Students should have their scores and levels tracked from grade to grade. Given that implementation of the Common Core is expected to strengthen year by year, percentages of students in this cohort scoring at or above Level 3 should either increase each year or at least hold true to the expected percentages established through vertical articulation in October 2014. Thus, the focus should be on deviations from these two possible patterns.
- *Consistency of cut scores across grades:* The results of the cross-grade review committee and subsequent actions by the Governing States provided a progression of cut scores rather than a single cut score. The reasonableness of the final distribution of cut scores and percentages of students at or above any given level can be compared to teachers' evaluations of their own students in the spring of 2015. This study will be carried out in conjunction with the teacher rating study in the spring of 2015.
- *Differential outcomes:* The Smarter Balanced assessment design was based on the principles of universal design. Therefore, one would expect equal access to all assessments and minimal differential item functioning (DIF). The DIF analyses proposed by Sireci (2012) for test development have been carried out. Similar analyses should be conducted in the spring of 2015.

### Validation Studies with External Variables

External variables are those outside of the tests themselves, in terms of how they relate to performance on the tests. These include teacher ratings, student grades, scores on other tests, employer ratings, and other variables. Specific examples are outlined below.

- *Teacher ratings in 2015:* Teachers in selected schools will provide ratings of samples of students, using the threshold ALDs. It will then be possible to cross-tabulate those ratings with Smarter Balanced test scores and level designations.
- *Teacher ratings in subsequent years:* If a Grade 5 student is deemed ready to move on to Grade 6 and perform adequately, the Grade 6 teacher should find that student ready. To the extent that faculty in subsequent years find students not to be as prepared as the level

designations they received the previous year, either the cut scores were invalid, or there was a mismatch between prior-year scholastic preparation and subsequent-year requirements; i.e., misalignment of curriculum and instruction. Starting in the fall of 2015, selected teachers in Smarter Balanced states will be asked to categorize their students, using the Smarter Balanced ALDs (for the previous grade). The level designations they assign to their incoming students will be compared to the level designations those students earned on the previous spring's tests. This study will be repeated in the fall of 2016 and fall of 2017 with different schools, teachers, and students.

- *Student grades in 2015:* Other selected schools will provide class grades or course grades of samples of students who also take the Smarter Balanced tests. It will then be possible to cross-tabulate those course grades with Smarter Balanced test level designations.
- *Course grades in subsequent years:* A parallel study will focus on course grades in 2016 and 2017. Students rated at the top level in Grade 6 should outperform students rated at lower levels when they are evaluated on Grade 7 work. College- and career-ready high school students should perform better in college algebra and freshman English than those who are not considered college and career ready. Starting in the 2015-16 school year, selected schools in Smarter Balanced states will be asked to supply course grades for their students. These grades will be compared to the level designations those students earned on the previous spring's tests. This study will be repeated in the fall of the 2016-17 and 2018-19 school years, by which time two cohorts of high school students will have entered postsecondary education and can supply college course grades.
- *NAEP scores:* Many of the schools testing in the spring of 2015 will also administer the National Assessment of Educational Progress (NAEP). For students taking those tests, level designations and/or scale scores may be available. In the event that they are, those scale scores and level designations can be compared to Smarter Balanced scale scores and level designations. This is essentially the approach taken by Gary Phillips (2012) in comparing NAEP scores to state achievement test scores.
- *Scores on other tests:* Many states, districts, and individual schools will continue to administer other standardized assessments, either commercial off-the-shelf tests or additional state-sponsored tests. Samples of students who take both a Smarter Balanced assessment and an additional standardized test in 2014-15 will provide the data for these studies. It will be necessary to obtain not only scale scores but also percentile ranks, proficiency levels, or other derived scores and scales for the external tests.
- *Scores of college students on the Grade 11 tests:* Although higher education faculty will be involved in the setting of achievement levels for the high school tests, there is no substitute for administration of the Grade 11 tests to college freshmen during the 2014-15 school years and cross-tabulation of level designation with their course grades. Similar studies have been conducted as part of the American Diploma Project (Miles, Beimers, & Way, 2010). Samples of college students, drawn from a variety of institution types, should take Smarter Balanced high school tests and also report their course grades in freshman English or mathematics. Similarly, samples of high school freshmen should take the grade 8 tests, and samples of grade 6 students should take the grade 5 tests.
- *Differential prediction:* Prediction of future outcomes is but a first step; comparing predictability across subgroups is the second step. In particular, it is advisable to compare the predictive power of Smarter Balanced assessments for students in general with their predictive power for specific target groups.
- *Opportunity to learn:* Curricular and instructional validation must be considered, especially over time. The first opportunity-to-learn (OTL) survey should be conducted in the spring of 2015 concurrent with an operational administration of Smarter Balanced tests. No matter

how well the tests are constructed, no matter how well they are aligned with the Common Core, and no matter how carefully cut scores are derived, if large numbers of students have not had the opportunity to learn the content of the tests, no cut score will be meaningful. Students in states adopting and implementing the Common Core early would be expected to perform better on Smarter Balanced tests than students in later adopting and implementing states. These studies should provide concrete evidence to support or dispel those expectations.

- Employer evaluations:* The study of course grades in subsequent years will cover the college half of “college and career ready.” Surveys of employers should cover the career half. During the 2014-15 school year, businesses and industries that hire large numbers of students right out of high school should be identified. Representatives of those businesses should identify minimum academic skill requirements of entry-level employees. By the 2016-17 school year, many of the high school students who took Smarter Balanced tests in 2014-15 will have entered the work force. In the fall of 2016 and again in the fall of 2017, selected employers of those students who have entered the work force in those years (having taken the grade 11 tests the previous year) should receive survey forms to complete regarding the readiness of those young people for the jobs they have taken. Responses from those employers would be matched with the Smarter Balanced scores and level designations of their employees.

**Organization and Implementation of Studies**

Table 15 shows a proposed organization and implementation timeline for the various studies outlined above. It is intentionally general in nature, prescribing neither sample size nor specific analytic tools or procedures.

Table 15. Validation Study Implementation Timeline.

School Year	Internal Validation Studies	External Validation Studies
2014-15	Interim assessments – collect data from selected schools and districts showing the relationships between interim assessments and subsequent summative assessments, particularly the relationship between the summative assessments and those interim assessments taken shortly before them.	Teacher ratings – collect teacher ratings from selected schools and districts, and compare their evaluations of student levels to those obtained on Smarter Balanced tests.
	Formative assessments – collect data from users and non-users of formative assessments to compare summative performance.	Student grades – collect student grades from selected schools and districts, and compare those grades to student scale scores and levels on Smarter Balanced tests.
	Digital library – collect data from users and non-users of the digital library to compare summative performance.	NAEP scores – identify students who will participate in the National Assessment and arrange to match their NAEP scores to their Smarter Balanced scores to solidify the link between the two scales.
	Expert studies – recruit higher education faculty and teachers from each grade to evaluate proto-tests or take a computer	Scores on other tests – identify schools and districts administering at least two commercially available tests and at least

	adaptive version of a specific test and provide feedback to Smarter Balanced with regard to alignment with the Common Core, rigor, and difficulty.	two state- or locally-administered tests; match student scores, and report correlations, scale equations, and differential scores on other tests by Smarter Balanced level.
	Consistency of cut scores – examine percentages of students at each grade level to determine whether the pattern of percentages in 2014 holds up.	Scores of college students on grade 11 tests – identify 4-year and 2-year colleges to participate; administer Smarter Balanced grade 11 ELA tests to selected students, and report distributions of scale scores and levels.
	Differential outcomes – perform DIF analyses on all items by race, gender, and program.	Opportunity to learn – identify a sample of districts and schools in all states administering Smarter Balanced states, and administer an OTL survey to teachers and administrators. Compare test results to OTL rates.
2015-16	Interim assessments – repeat 2014-15 study. Report cumulative results as well as trends.	Teacher ratings – collect teacher ratings from receiving teachers and compare to level designations from 2015 Smarter Balanced assessments.
	Formative assessments – repeat 2014-15 study. Report cumulative results as well as trends.	Student grades – compare course grades of students in each grade to their previous year’s Smarter Balanced level designation. Tabulate average grades by Smarter Balanced level.
	Digital library – repeat 2014-15 study. Report cumulative results as well as trends.	Differential prediction – carry out studies of teacher ratings and student grades by subgroup (race, gender other reporting categories).
	Level on subsequent tests – for selected districts and schools, merge 2015 and 2016 assessment data and compare level designations. Report percentages of students moving up, down, or remaining in the same level.	Opportunity to learn - - repeat 2014-15 study. Report cumulative results as well as trends.
	Consistency of cut scores across grades – repeat 2014-15 study. Report cumulative results as well as trends.	Employer evaluations – for students tested in grade 11 in 2015, collect employer evaluations of new 2016 graduates by October 2016.
	Differential outcomes - repeat 2014-15 study. Report cumulative results as well as trends.	

	Differential outcomes- repeat 2014-15 study. Report cumulative results as well as trends.	
2016-17	Interim assessments – repeat 2015-16 study. Report cumulative results as well as trends.	Teacher ratings – repeat 2015-16 study. Report cumulative results as well as trends.
	Formative assessments – repeat 2015-16 study. Report cumulative results as well as trends.	Student grades – repeat 2015-16 study. Report cumulative results as well as trends.
	Digital library – repeat 2015-16 study. Report cumulative results as well as trends.	Differential prediction – repeat 2015-16 study. Report cumulative results as well as trends.
	Level on subsequent tests – for selected districts and schools, merge 2015 and 2016 assessment data and compare level designations. Report percentages of students moving up, down, or remaining in the same level.	Opportunity to learn – repeat 2015-16 study. Report cumulative results as well as trends.
	Consistency of cut scores across grades – repeat 2015-16 study. Report cumulative results as well as trends.	Employer evaluations – for students tested in grade 11 in 2016, collect employer evaluations of new 2017 graduates by October 2017.
	Differential outcomes - repeat 2015-16 study. Report cumulative results as well as trends.	
	Differential outcomes- repeat 2015-16 study. Report cumulative results as well as trends.	
	Convene K-12 educators, higher education faculty and administrators, general public, and other key stakeholders to review cut scores set in 2014 in light of validation data collected to date. Recommend changes if necessary.	

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Author.
- Cizek, G. J. & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.
- Egan, K. A., Schneider, C., & Ferrara, S. (2012). Performance level descriptors. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2<sup>nd</sup> Ed.), New York: Routledge.
- International Test Commission (2010). *Guidelines for translating and adapting tests*. Downloaded from the world wide web at <http://www.intestcom.org> on September 1, 2014.
- Johnstone, C. J.; Altman, J.; & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available from <http://www.cehd.umn.edu/nceo/OnlinePubs/StateGuideUD/default.htm>
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational Measurement* (4<sup>th</sup> ed., pp. 17-64). Washington, DC: American Council on Education/Praeger.
- Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schultz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations*. New York: Routledge.
- Lewis, D. M., & Mitzel, H. C. (1995). An item response theory based standard setting procedure. In D. C. Green (Chair), Some uses of item response theory in standard setting setting. Symposium conducted at the annual meeting of the California Educational Research Association, Lake Tahoe, NV.
- Lewis D.M., Mitzel, H. C., Green, D. R. (1996). Standard Setting: A Bookmark Approach. In D. R. Green (Chair), IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring. Symposium presented at the 1996 Council of Chief State School Officers 1996 National Conference on Large Scale Assessment, Phoenix, AZ.
- Loomis, S. C. (2011). *Toward a validity framework for reporting preparedness of 12<sup>th</sup> graders for college-level course placement and entry to job training programs*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
- Miles, J. A., Beimers, J. N., & Way, W. D. (2010). The Modified Briefing Book Standard Setting Method: Using Validity Data as a Basis for Setting Cut Scores. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Denver, CO.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Phillips, G. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations*. New York: Routledge.

- Sireci, S. G. (2012). Smarter Balanced Assessment Consortium: Comprehensive Research Agenda. Report Prepared for the Smarter Balanced Assessment Consortium.
- Smarter Balanced Assessment Consortium (2010). *Theory of Action: An Excerpt from the Smarter Balanced Race to the Top Application*. Tacoma, WA: Author.
- U.S. Department of Education (2009, January). Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001. [Revised December 21, 2007 to include Modified academic achievement standards. Revised with technical edits, January 12, 2009] Washington, DC: Author.

## Appendices

[Appendices are available upon request.]