Observation and Report on

Smarter Balanced Standard Setting

October 12–20, 2014

Prepared by:

Gregory J. Cizek, PhD
and
Heather Koons, PhD

October 23, 2014

I. Overview

The Smarter Balanced Assessment Consortium (hereafter *Smarter Balanced*) desired to establish achievement levels on newly developed assessments in English Language Arts (ELA) and Mathematics. Large and representative groups of qualified participants were empaneled to perform the standard setting procedures for high school assessments (11th grade) on October 13-15, 2014, for middle grades (6-8) assessments on October 15-17, 2014, and for elementary (grades 3-5) assessments on October 17-19, 2014. The three two and one-half day sessions for each test were implemented using an adaptation of the Bookmark procedure (Lewis, Mitzel, Mercado, & Schulz, 2012). The in-person sessions were preceded by an on-line input gathering process and followed by a vertical articulation process comprising a stakeholder review of the performance standards recommended by the participants. The standard setting activities were conducted by contractors Measurement, Incorporated (MI) and CTB/McGraw-Hill (CTB).

The authors of this report responded to a request by Smarter Balanced that an independent, external auditor be present for the sessions to verify that the standard setting process and procedures followed the plan that had previously been reviewed by the Smarter Balanced Technical Advisory Committee and approved by Smarter Balanced governing states, and to submit a report of observations and findings. The auditors selected, Dr. Gregory J. Cizek and Dr. Heather H. Koons, have expertise and extensive experience in the area of setting performance standards (see, e.g., Cizek, 2001, 2012; Cizek & Bunch, 2007; Cizek, Bunch, & Koons, 2004). Dr. Cizek was the lead auditor and was present to perform observations of the in-person sessions on October 13-16 and 18-19 and the vertical articulation session on October 20th. Because of a prior commitment, Dr. Cizek was unavailable to observe on October 17th; Dr. Koons observed on that day. Drs. Cizek and Koons coordinated their observations to include pre-workshop

communications, an on-site briefing when Dr. Koons arrived at the sessions; a short period of overlapping observations on October 16th; a second debriefing on the morning of October 18th; and an additional time of overlapping observations on October 18th. Both auditors contributed to the preparation of this report summarizing their observations, and both reviewed interim report drafts. The conclusions expressed in the Summary and Recommendations section of this report are those of Dr. Cizek, in consultation with Dr. Koons.

This report provides a description of the standard setting activities, some recommendations, and a summary evaluation. The report is organized into five sections: 1) Overview; 2) On-line Procedures; 3) In-person Observations; 4) Vertical Articulation; 5) Summary and Recommendations; and 6) References.

II. On-line Procedures

Demonstrations of software used for the on-line procedures (and to be used in the operational standard setting), and software used for automated assembly of the Ordered-Item Booklets (OIBs) were provided at a meeting on October 7, 2014 at the MI headquarters in Durham, North Carolina. The meeting provided information on the materials made available to the on-line participants, which included a 12-page document on how to use the on-line tool; links to various Smarter Balanced assessment resources, and a practice activity. A field test of the on-line materials and procedures was conducted by MI in Durham, North Carolina in August 2014.

The on-line component of the standard setting was opened for participation on Monday, October 6, 2014 and the last window closed on Saturday, October 18, 2014 at 11:59 p.m. Eastern time. Participants using the on-line tool were able to review an OIB and provide their judgment regarding placement of a single bookmark for Level 3. A practice activity and various resources were provided for participants. Among the reference materials available to participants were the achievement level descriptors (ALDs), the content specifications for the test, bookmark placement instructions, and relevant practice tests. Participants were then provided with a six-item orientation round OIB which was provided to help familiarize them with the bookmarking task. For each item in the OIB, certain information was provided to participants, including target and claim identification, Depth of Knowledge (DOK) addressed, stimulus materials, and key/rubric/scoring information.

During the practice activity, set one bookmark for Level 3. It was possible, however, for participants to have skipped completing the practice OIB, although the on-line tool recorded if the practice activity was performed; it was not recorded if participants reviewed the ALDs, if they reviewed the on-line tool information provided, or if they took practice test before actually engaging in the OIB review and providing their judgments. Participants then completed the

operational OIB, comprising approximately 70 items, and again placed one (Level 3) bookmark. Participants did not need to review every item; they could skip to a portion of OIB where they believed the Level 3 bookmark would most likely be appropriate. Some NAEP and PISA items were embedded in the OIBs and identified as such for participants; of the NAEP and PISA items viewed, at least on NAEP item (Grade 11 ELA) appeared to be missing a stimulus, and a PISA item included spellings that might not be familiar to students accustomed to North American English (e.g., “labour").

During their OIB review, participants were permitted to take a break at any time. At the end of the activity, participants who submitted their ratings were offered the opportunity to receive a certificate of completion; the certificate was automatically provided to any participant who completed their bookmark placements using three hours or more; however, there was some confusion about the certificates with some participants who took less than three hours being asked about their desire for a certificate, but who were not provided with one.

Overall, the on-line tool appeared to be well-designed and appropriate for the task of collecting on-line judgments. For the future, or as research activities for the present standard setting, some suggestions would include:

1) It may be desirable to produce an FD record of how many items were reviewed by on-line participants.
2) It may be desirable to improve or simplify on-screen navigation. Performance tasks with stimuli seemed somewhat difficult to navigate between several screens, especially considering the need to view the item, the stimuli/passages, scoring guides, ALDs, exemplar responses for constructed-response items, and so on.
3) It would also seem desirable to provide participants with scoring rules for all multi-part

items (e.g., Must all be correct to get credit? Is each element scored separately?).

4) It may be desirable to keep the text with explicit directions for the bookmarking task on screen at all times when actual item judgments are being made.

In addition to review of the on-line tool, an interview and demonstration was conducted with the creators of OIBs. An automated assembly program was used to create the 14 OIBs needed for the operational standard setting. Overall, it appeared that the procedures were very successful in creating OIBs that appropriately covered the test specifications.

III. In-person Observations

Procedures for the in-person standard setting workshops and vertical articulation were piloted during the week of August 18, 2014 in Durham, NC. Dr. Koons was present to observe the piloting from from 1:30 p.m. through the end of the day on August 18th and from 1:30-3:30 p.m.on August 19th. Dr. Koons provided feedback to Dr. Bunch of Measurement Inc., based on her observations.

Prior to the operational sessions, the auditors were provided with several documents for review. These materials included:

* Smarter Balanced Achievement Level Setting Plan
* Smarter Balanced Achievement Level Setting Plan Supplement
* Selecting Items for the ALS
* Achievement Level Setting Agenda
* In-person Standard Setting Facilitator Scripts
* In-person Workshop PowerPoint Slides

The operational sessions took place from October 12-19, 2014 in Dallas, Texas. Three two and one-half days sessions were conducted for ELA and mathematics: one for high school (grade 11); one for middle grades (6-8); and one for elementary grades (grades 3-5).

The first sessions (grade 11) were held on October 13-15, 2014. The evening prior to this session, a brief orientation was provided for selected participants who were identified to serve in the role of table leaders. The following morning, the operational standard setting began. A target sample of 72 panelists for each subject area was desired; however some participants cancelled plans to attend the workshop in the days immediately preceding the event. The grade 11 participants (as well as participants on other grade level panels) were chosen to be representative

of the teaching population of Smarter Balanced member states; final demographic characteristics of all participants will be documented in a forthcoming technical report produced by the standard setting contractor. For grade 11, a total of 68 participants in ELA and 70 participants for mathematics were empaneled in four rooms as follows:

* 34 ELA panelists in Room A (ELA-A)
* 34 ELA panelists in Room B (ELA-B)
* 35 Mathematics panelists in Room A (Math-A)
* 35 Mathematics panelists in Room B (Math-B)

For grades 6, 7, and 8, a total of 87 participants in ELA and 89 participants for mathematics were empaneled in six rooms as follows:

* 30 Grade 6 ELA panelists
* 27 Grade 7 ELA panelists
* 30 Grade 8 ELA panelists
* 30 Grade 6 Mathematics panelists
* 30 Grade 7 Mathematics panelists
* 29 Grade 8 Mathematics panelists

For grades 3, 4, and 5, a total of 80 participants in ELA and 88 participants for mathematics were empaneled in six rooms as follows:

* 26 Grade 3 ELA panelists
* 27 Grade 4 ELA panelists
* 27 Grade 5 ELA panelists
* 30 Grade 3 Mathematics panelists

* 29 Grade 4 Mathematics panelists

* 29 Grade 5 Mathematics panelists

The following sections provide the auditors' observations for each of the grade band sessions. One caveat should be noted. At all times, one auditor was present to view the activities. However, at any given time, there were four or six breakout rooms working on the standard setting tasks. It was not possible to view all rooms at the same time, nor did it seem advisable to observe only one of the rooms for an entire session. Instead, auditors attempted to spend smaller blocks of time in each room. The advantage of this strategy was that the auditors gained a fairly good sense the approaches and particularities of each group. A disadvantage of this strategy was that the activities of a single group were not typically observed from start to finish. Thus, details of how sessions progressed were sometimes not fully apprehended until the observations were pieced together across several days and sessions. As a consequence of this, the details of some aspects of the sessions are dispersed throughout the report. That is, some aspects of the sessions were able to be fully reported on Day 1 of a panel's activities, some on Day 2, and some on Day 3; some aspects were fully observed in one grade level band, other aspects were fully observed in another grade level band. Readers are encouraged to consider the entire report to gain a complete picture of the entire process.

Grade 11 Sessions, Day 1: Morning Activities (Monday, October 13, 2014)

Grade 11 session activities began with participant registration and breakfast the morning of the first day. All registration and other logistics appeared to proceed without any issues. After breakfast, participants met in a full group (i.e., ELA and mathematics, combined) for an orientation session in a separate ballroom. The whole group orientation session occurred from 8:30-9:30 a.m. on the first day. It began with a welcome and overview presentation by Dr. Joe

Willhoft, Executive Director of Smarter Balanced. The presentation included an overview of the general purpose of the sessions, provision of some background and context, questions, solicitation of participants' concerns (e.g., role of Smarter Balanced at college entry; role of Smarter Balanced in career training readiness; test taker need for keyboarding skills), the introduction of key Smarter Balanced content leadership in ELA and mathematics and announcement of the presence of state and other observers. Finally, a "parking lot" was announced, whereby participants could indicate concerns or questions to their individual room facilitators, with those concerns and questions to be addressed subsequently in a whole group session.

The orientation continued with a presentation by Dr. Mike Bunch, Senior Vice President of MI, who described the agenda and goals for the workshop, along with brief attention to background on the common core state standards (CCSS), the ALDs, the practice test, the bookmark procedure, the concept of an OIB, performance tasks, evaluations, and "key shifts" embodied in the CCSS such as less breadth but deeper, more conceptual understanding, emphasis on cross grade coherence and application in mathematics, and more complex texts and increased presence of nonfiction texts in ELA. Dr. Bunch reminded participants of the focus on claims across subjects, with the four ELA claims focusing on 1) Reading, 2) Writing, 3) Speaking/Listening, and 4) Research/Inquiry; and mathematics claims focusing on 1) Concepts and Procedures, 2) Problem Solving, 3) Modeling and Data Analysis, and 4) Communication and Reasoning. The presentation introduced participants to the four types of ALDs—Policy, Range, Threshold, and Reporting—and indicated that the focus in the workshop would be on the threshold ALDs describing the knowledge and skills of students just entering Levels 2, 3, and 4. The presentation also included a brief overview of the question and test development history and processes, the various item formats comprising Smarter Balanced assessments, including selected-response (SR), constructed-response (CR), and technology enhanced (TE) formats, as well as information

on the mode of administration (computer-based), and available supports and accommodations for SWDs. Finally, Dr. Bunch noted a few "ground rules" for the workshop, including the fact that they were assembled to recommend, not set, performance levels, the requirement that participants sign a security/confidentiality agreement, and that participants engage in a group process where each panelist can freely contribute to discussions.

Dr. Bunch then introduced Deborah Sigman, co-chair of the Smarter Balanced executive committee, who thanked participants for their involvement and their efforts with Smarter Balanced to improve teaching and learning for students. To conclude the session, Dr. Bunch provided information on the room assignments for the mathematics and ELA groups and introduced the room facilitators:

HS ELA-A: Craig Deville (MI) and Anne Wilder (MI)

HS ELA-B: Gretchen Shultz (CTB) and Ric Mercado (CTB)

HS Math-A: Winne Reid (MI) and Maude Eno (MI)

HS Math-B: Judy Hickman (CTB) and Jennifer Lord-Besson (CTB)

Participants were then dismissed to their content area breakout rooms. At the conclusion of the session, Dr. Willhoft asked observers to remain in the ballroom, where he conducted a brief session to instruct the state-representative observers in the ground rules for their observations. (Smarter Balanced member states had been afforded the opportunity to send observers to the in-person standard setting activities.). Dr. Willhoft welcomed the representatives and admonised them to engage in non-intrusive movement between rooms, to avoid engaging in conversations with panelists while in the session rooms (although OK to do so at breaks or afterward), prohibiting them from speaking after the standard setting sessions about the results of the sessions, about specific content on the assessment, or about individual panelists, but encouraging them to

feel free to speak freely after the sessions with colleagues and others about the standard setting process.

Participants began working in their breakout groups beginning at 9:45 a.m.. Rooms were arranged in rectangular tables of six panelists, with each panelist having his/her own computer monitor, keyboard, and (for ELA) headsets. Participants were instructed to stay at the same table for the session, but they could change seats within a table group if they desired. In general, the meeting spaces seemed somewhat cramped, with little work surface for participants, tables that were fairly close to each other, and height of monitors that sometimes made it difficult for participants on opposite sides of a table to engage in discussion. (Some participants could more easily communicate with persons next to them or at an adjacent table than with other participants at their own tables.) Overall, however, the arrangements did not appear to present an overwhelming obstacle to the conduct of the sessions and seemed conducive for participants to perform their tasks.

The breakout sessions began with a welcome from the facilitators, individual panelist introductions, housekeeping information (such as non-disclosure forms, multimedia permission release forms (for audio, video, and photographic recording of participants), and a panelist information questionnaire. An introduction was then provided to the software to be used in the sessions. The software appeared to work well; isolated minor issues were addressed quickly and effectively by MI personnel. Facilitators then reinforced and extended information on the claims, targets and standards covered by the Smarter Balanced assessments, and provided additional information on the ALDs, with a focus on the threshold ALDs. Participants were reminded that they would be setting a single bookmark for Level 3 during an orientation session, and they would be setting three bookmarks during the operational bookmarking, beginning with the Level 3 bookmark, followed by Level 4, then Level 2. The rooms each adjourned for a break at

approximately 10:40 a.m., and resumed at approximately 10:55 a.m.

When they reconvened, the groups engaged in several tasks to prepare them for operational standard setting, including taking the Smarter Balanced "training test" that allowed them time to explore SR, CR/performance tasks, and TE item formats, experience with the keyboarding/mousing activities that students would need to apply, and experience with the sound and headphones used (for ELA). The review of on-line and technology features appeared to go smoothly in each room; the monitors appeared to be large and clear enough for ease of use, with stimuli and items both appearing on the same screen and limited scrolling required. Participants then reviewed the Smarter Balanced "practice test." It was somewhat unclear how the purposes of the "training tests" and "practice tests" differed, although the practice test appeared to give participants an opportunity to review some of the available accommodations for SWD (e.g., text to speech).

At approximately noon on Day 1, a room-level activity was conducted, with participants solicited for their reactions to the item reviews. Participants were generally favorable in their perceptions, noting that the items and tasks appeared to reflect rigorous, real-life, and transferable knowledge and skills. Some less favorable observations touched on the limitations of the test interface that did not allow cut-and-pasting from source documents, the cognitive demand of "topic jumping" from one challenging task to another within the assessment, and the general impression that the rigor of the assessments was not well aligned to the current abilities of many students and the lack of instructional resources available in schools to support instruction covering the CCSS. The groups then broke for lunch at approximately 12:25 p.m., with instruction to reconvene in the ballroom for a whole group activity at 1:20 p.m.

Grade 11 Sessions, Day 1: Afternoon Activities (Monday, October 13, 2014)

At the end of the lunch break on Day 1, in a whole group session, Dr. Willhoft and Mr.

Tony Alpert, Chief Operating Officer for Smarter Balanced, read and answered "parking lot questions" that had arisen in each of the rooms. The lunch break/parking lot activity ended at 1:20 p.m., and participants transitioned to the ballroom for a whole-group activity.

At approximately 1:35 p.m., Dr. Dan Lewis (CTB) initiated a whole-group presentation in the ballroom. The presentation began with an overview of remainder of the day and the processes in which participants would engage, including an introduction to the interface that panelists would be using, the resources, keys/scoring guides/rubrics that would be available, the OIB, general information on the bookmarking process, and specific directions on the cognitive analysis of each item they would be performing. Dr. Lewis provided specific information on the two questions participants must consider when engaging in the cognitive analysis: 1) What knowledge and skills must a student have in order to know the correct answer to an item? and 2) What characteristics of the item make it more challenging than the preceding item(s). Although the slides used by Dr. Lewis were produced using a small font that was somewhat difficult for the participants to see, the explanations were clear and participants appeared to grasp the key aspects of the presentation.

Participants were then dismissed to return to their content area rooms. From approximately 2:00 p.m. through the end of Day 1, participants worked in the breakout rooms on the task of reviewing their item map/OIBs and gaining an integrated conceptualization of the content of the Smarter Balanced assessment. Using the on-line interface, participants worked through the OIB, answering the two main questions, and adding comments on items to note the knowledge and cognitive skills tapped by each items. This activity continued throughout the afternoon, although many participants did not appear to have sufficient time to complete the activity by the end of the day.

At 2:15 p.m. on Day 1, CTB personnel provided one of the auditors (Cizek) with a demonstration of Bookmark Pro software—the software that would be used to derive performance

standards for all of the Smarter Balanced assessments under consideration. The software actually performs multiple functions, including the collection of panelist judgments, data analysis, and production of graphic and tabular results at the end of each round. Overall, the software appeared to be a very streamlined and useful utility for standard setting.

All participants were dismissed for the day between approximately 5:00 and 5:30 pm on Day 1. At approximately 5:30 p.m., a debriefing session was conducted in the ballroom, attended by contractor staff, Smarter Balanced representatives, and the auditor. Two major issues were raised: 1) some software concerns that arose early on Day 1, and 2) the insufficient time allocated for participants to complete the cognitive analyses of the OIBs. With respect to the first issue, facilitators noted that they had made slight adjustments in their presentations, such that the issue did not appear to be consequential. A considerable amount of time was spent discussing the second issue. It was decided to alter the Day 2 schedule to allow time for participants to complete the Day 1 activity in the morning by inserting some "catch up" time and slightly shortening the time allocation for Round 1 bookmark placements.

Grade 11 Sessions, Day 2: Morning Activities (October 14, 2014)

The second day of grade 11 standard setting consisted of a whole-group opening session beginning at 8:30 a.m. in the ballroom. The session commenced with Dr. Bunch announcing the adjustments in the agenda to allow participants to complete their cognitive analyses of the OIBs. The next portion of the session consisted of a presentation by Dr. Lewis who provided a more detailed description of the bookmark procedure and specific information and instructions on threshold students, bookmark placement for Level 3, Level 2, and Level 4. Dr. Lewis then answered questions from the total group and offered clarifications on issues such as: Why are some items repeated in the OIB? What about performance of threshold students on the items after the

bookmark? What is the relationship between threshold performance and specific targets? What about items that appear later in the OIB, but that appear to assess very fundamental skills? Is Level 3 passing? The whole group session ended at approximately 9:00 a.m. and participants were dismissed to go to their breakout rooms.

From 9:00-9:45 a.m., participants engaged in the "catch up" activity to complete the cognitive analysis of the OIBs. By approximately 9:45 a.m., all groups had begun a practice activity in which they placed a Level 3 (only) bookmark using a six-item orientation-round OIB and other resources. During this activity, one ELA room took early break (10:05 a.m.) because of some connection issues experienced by participants who were not able to continue the practice activity. (Apparently, the connectivity issue was not limited to one room or to the practice activity. However, the on-site IT personnel worked to resolve issues each time they arose and issues appeared to decline as the Day 2 activities continued.) The group reconvened at approximately 10:20 a.m. and all participants finished the practice ratings. Ratings were then collected and facilitators polled the room to see where bookmarks had been set. Software for presenting results to the group projected on a screen did not appear to work, so medians, etc. could not be presented. (The software worked correctly in another room observed.) The facilitator addressed the issue by asking for a show of hands for bookmark placements on each page. Discussion ensued as to participants' rationales for bookmark placements on selected pages.

Operational Round 1 bookmark placements then began at approximately 10:45 a.m. In one room observed, panelists were directed to place their bookmarks independently, first for Level 3, then Level 4, then Level 2; in another room observed, panelists were told that they could work together to place their Round 1 bookmarks. In all rooms observed, panelists were also directed to complete the first questionnaire upon closing out of the practice round before beginning their Round 1 operational bookmark placements.

In at least two of the rooms observed, participants appeared to have continuing, sporadic difficulty with connectivity of their computers. Some participants completed their task beginning at approximately 11:30 a.m. and were instructed to return for the next activity following lunch at 12:50 p.m. Subsequently, participants were instructed to return at 2:00 p.m.

Grade 11 Sessions, Day 2: Afternoon Activities (October 14, 2014)

The Day 2 afternoon session began at 2:00 p.m., with participants meeting in breakout groups for an introduction to the process that would be used to discuss the Round 1 results and to place the Round 2 bookmarks. Some technological difficulties were again experienced; however, facilitators were still able to present relevant information on the Round 1 bookmark placements for the group to consider, and technical specialists were sought out to help remedy the problems. A block of time was then allocated for participants to discuss the rationales for their Round 1 bookmark placements in table groups, beginning with their placement of the Level 3 bookmark. (The same process was used for Grade 11 and subsequent grade level sessions; namely, discussions following the setting of Round 1 bookmarks were at the table level; discussions following Round 2 were at the room level.) At the conclusion of the discussion, participants were dismissed for a short afternoon break at approximately 3:10 p.m.

The groups reconvened at approximately 3:25 p.m. They were presented with information on the on-line panel's results and an end-of-round questionnaire. At one point during the afternoon activities, Dr. Willhoft and Mr. Alpert visited the panel rooms to answer specific content area questions that were more relevant to the breakout groups than to the whole group and were not addressed during the lunch-hour "parking lot" presentation. The groups then were instructed to finish completing the end-of-round 1 questionnaire, submit the questionnaire, and the Round 2 activities began.

To complete the Round 2 bookmark placements, the groups were instructed to go back to their item maps. The item maps used for Round 2 contained new information on item scale locations. The scale location information allowed participants to discern relative differences in OIB item difficulties so that, if electing to move their bookmarks for Round 2, the relative magnitude of impact of a move on the resulting cut scores could be seen. New information on NAEP and PISA item difficulty was also presented on the item map used for Round 2. Facilitators attempted to describe how the NAEP and PISA performance should be taken into account, although some participants seemed unclear as to precisely how to use the information. To complete their Round 2 ratings, participants were instructed to again begin with a review of their Level 3 bookmark placement, then proceed to review their Level 4 bookmark, then their Level 2 bookmark. In two of the groups, some participants appeared to complete their Round 2 bookmark placements by approximately 4:15 p.m. with nearly all participants completing their work by 4:30 p.m. In the other two rooms, most participants did not finish until approximately 5:15 p.m. or later. At the end of Day 2, all participants were instructed to complete the Round 2 questionnaire and were dismissed individually for the day, noting that the activities for Day 3 would be begin at 8:30 a.m. in grade/content area breakout groups, not as a whole group as had been the case for the first two days.

At the conclusion of Day 2, another debriefing session took place at 5:30 p.m.in the ballroom, involving contractor staff, Smarter Balanced representatives, and the auditor. No major issues from the day were identified by the group.

Grade 11 Sessions, Day 3: Morning Activities (October 15, 2014)

The final day of grade 11 standard setting began at 8:30 a.m. with subject area groups meeting in their breakout rooms to begin Round 3. The sessions began with the facilitator

providing information related to logistics (reimbursement forms, airport shuttles, etc.), followed by some questions and answers about the process to date and what would be the next steps for the results.

The facilitator then presented a graphic display of participants' page judgments for the three achievement levels. A tabular breakdown of results (median bookmark placements) by table and room overall was then presented. Another slide was then shown to participants, providing them with impact data, based on their median Round 2 recommendations, and based on Smarter Balanced assessment field test data. A discussion ensued regarding the group's perspectives on the impact, what the distribution of student performance should look like, and other issues.

In one of the rooms observed (an ELA panel), the distributions of page judgments across the levels appeared to be fairly unimodal and with modest variability. The group appeared in general to find the impact to be realistic, with the possible exception of the Level 2 results. Some panelists noted that they put forth greatest effort in their work on the Level 3 and Level 4 cut scores, experiencing some fatigue by the time they reached the Level 2 cut, combined with some sense that the Level 2 threshold description was somewhat less clear for them in terms of operational utility. In another group (Math), there seemed to be polarization in the results, which were uniformly bimodal across the three levels. The discussion in this group appeared to reveal a lack of clarity about the appropriate referent for the standard setting task, with one viewpoint (from those in the higher modal area of the distributions) being that the content of the items accurately represented what the threshold description indicated, and the other viewpoint (from those in the lower modal area of the distributions) being that their judgments reflected more appropriate expectations for what "real students" can do.

Observation of one other room suggested what might be a moderate concern: Although in earlier rounds, there appeared to be an understanding among panelists about grounding judgments

exclusively in the threshold descriptions, as more impact and other data were provided to participants, the more some appeared to "drift" from grounding their judgments exclusively in the threshold and range ALDs. More than one panelist articulated what were clearly wrong strategies for how they placed their bookmarks: one was wrong in the sense that they were not explicitly attempting to translate the threshold descriptions into page judgments based on the content of the items; another was wrong in the sense that they placed their bookmark at the first item where they thought the threshold Level N student fell below the RP50 criterion, as opposed to going beyond that item to see if other items beyond the identified item were still within the RP for the Level N student. It is possible that the drift might, to some extent, be attributable to a facilitator effect. (Over the course of the sessions, it seemed clear that facilitators in different rooms had differing degrees of experience and comfort leading standard setting workshops, an observation which will be elaborated upon later in this report.) One strategy for dealing with this drift might be to keep the bookmarking task posted at all times when panelists are making judgments. For example, during the cognitive analysis of the OIB activity, a slide with the two questions to be answered was kept on the screen throughout the entire activity so that panelists could constantly refer to it. Similarly, posting the precise bookmarking task directions might also help; for example:

1) "Referring to the threshold Level N student, place your bookmark on the first page where the threshold Level N student would have less than a 50% chance of answering the item correctly (or obtaining that score point or greater)."
2) "Don't stop at the first item where the chance of answering correctly drops below 50%, but examine items following the one you identify to be sure you have identified an appropriate bookmark location."

In all groups, an additional slide was then projected, showing participants the impact data

that would result, based only on the on-line participants' judgments. A final slide showed results separately for ethic groups, gender, LEP status, and other subgroups.

Participants then completed a questionnaire and began their Round 3 review and judgments. Before beginning their Round 3 judgments, the facilitators highlighted a new piece of information on the item maps: information on the range of items on the OIB item maps related to an ACT benchmark score. Discussion of the relevance of the ACT information ensued, and many participants questioned the relevance, representativeness, and usefulness of the ACT data. Participants' work on Round 3 began at approximately 9:30 a.m. Panelists took a break of approximately 40 minutes while the data from Round 3 were submitted for analysis and prepared for presentation. Three slides were shown to panelists: one showing the final median bookmark placements for the room, one showing the percentages (impact) that would fall into each category implied by the final group median placements, and one showing the percentages as a histogram. Panelists then completed a final questionnaire. The facilitators thanked the panelists for their work, provided additional information about lunch, airport transportation, reimbursement forms, and other logistics. As they were adjourned, panelists' materials were collected.

Grades 6-8 Sessions, Day 1: Morning Activities (October 15, 2014)

At the same time that the Grade 11 groups were completing their Round 3 ratings, the standard setting session activities for Grades 6, 7, and 8 ELA and mathematics began. The agenda for grades 6-8 mirrored that of grade 11, with table leader orientation occurring the evening of October 14, 2914 and general participant registration and breakfast the morning of October 15, 2014. After breakfast, grade 6-8 participants met in a full group (i.e., ELA and Mathematics, combined) for an orientation session in a separate ballroom. The whole group orientation session began with a welcome and overview presentation by Dr. Willhoft including an overview of the

general purpose of the sessions, provision of some background and context, questions, solicitation of participants' concerns. The orientation continued with a presentation by Dr. Bunch, who described the agenda and goals for the workshop, along with brief attention to background on the common core state standards (CCSS) including claims and targets, the item and test development process, assessment item formats, the ALDs, the practice test, the bookmark procedure, the concept of an OIB, performance tasks, evaluations, "key shifts" embodied in the CCSS, and the four types of ALDs with an emphasis on the threshold descriptions. Finally, Dr. Bunch noted the "ground rules" for the workshop, including the fact that participants were assembled to recommend, not set, performance levels, the requirement that participants sign a security/confidentiality agreement, and that participants engage in a group process where each panelist can freely contribute to discussions. To conclude the session, Dr. Bunch provided information on the room assignments for the mathematics and ELA groups, and introduced the room facilitators:

Grade 6 ELA - Chris Dunbar and Corey Palermo (MI)

Grade 6 Math - Kelly Bolton and Rick Mercado (CTB)

Grade 7 ELA - Molly Buck and Sarah Hagge (CTB)

Grade 7 Math - Lisa Johnson and Jennie Bowen (MI)

Grade 8 ELA - Amy Griswold and Dan Bowen (MI)

Grade 8 Math - John Upchurch and Jennifer Lord-Bessen (CTB)

Participants were then dismissed to their content area rooms where they began working in their breakout groups. The rectangular arrangement of tables that had been used for the grade 11 sessions was maintained for the the grades 6-8 sessions. One less table was needed per room, however, as only 30 panelists had been invited to participate for each grade/subject combination;

this fact permitted somewhat greater working space within the rooms and somewhat greater distance between tables. The set-up of the monitors and keyboards remained the same—an arrangement that posed a minor difficulty for some participants in terms of their ability to interact with other panelists seated across from them at their tables.

The breakout sessions again began with a welcome from the facilitators, individual panelist introductions, housekeeping information (i.e., non-disclosure forms, multimedia permission release forms (for audio, video, and photographic recording of participants), and a panelist information questionnaire. An introduction was then provided to the software to be used in the sessions. The software appeared to work well, although some intermittent connectivity issues were still noted. All participants again completed a training test and practice test. The groups then broke for lunch at approximately 12:30 p.m., with instruction to reconvene in the ballroom for a whole group activity at 1:20 p.m. The "parking lot" activity held previously at the end of the lunch hour was moved to more grade/content specific answers to questions presented by Dr. Wilhoft and Mr. Alpert who went to individual grade/subject rooms and addressed questions relevant to those grades and subjects.

Grades 6-8 Sessions, Day 1: Afternoon Activities (October 15, 2014)

The whole group activity to begin the afternoon consisted of the same presentation on the afternoon of the first day for the grade 11 panelists: Dr. Lewis presented an overview of the item map and the tasks that participants would be performing in the afternoon, with an emphasis on answering the two main questions about each item (i.e., what the item measures/sources of challenge, and what makes an item in the OIB more challenging than the preceding items. A change implemented was the inclusion of an additional aid and some explication of the specific bookmarking task in the presentation by Dr. Lewis. The slides projected in each room were also

modified to include "Bookmark Placement Instructions" that clearly stated the bookmarking task:

> "Ask yourself: Would a student at the threshold have at least a 50% chance of earning this point? If YES, move on to the next item. If NO, place your bookmark here."

Panelists were then dismissed to their breakout rooms and engaged in the cognitive analysis of the OIB for the remainder of the afternoon. Some grade/content groups finished the task by the end of the day; others were still completing the task when their groups were dismissed for the day at approximately 5:15 p.m.

As was done at the end of the first two days of the grade 11 sessions, a debriefing session was held with contractor staff, Smarter Balanced representatives, and the auditor in attendance. No major concerns were identified in the session. Some facilitators again observed that some members of their grade/content groups needed for additional time to complete their analysis of the OIB. To address that issue, the schedule for grades 6-8 Day 2 was again adjusted as it was for grade 11 sessions to add 45 minutes at the beginning of the next day for participants to complete that task. In addition, some facilitators noted intermittent connectivity issues, although the issue did not appear to impede any of the panelists' work. Finally, Dr. Lewis indicated his desire to include an additional slide in his opening Day 2 presentation to clarify how the term "consistently" in ALDs should be operationalized when panelist are making judgments about items using RP50. Discussion of the idea was generally favorable; some suggestions were made as to how the clarification should be presented, and the addition of the new slide was approved by the group.

Grades 6-8 Sessions, Day 2: Morning Activities (October 16, 2014)

The second day of grade 6-8 standard setting consisted of a whole-group opening session beginning at 8:30 a.m. in the ballroom. The session again commenced with Dr. Bunch announcing

the adjustments in the agenda to allow participants to complete their cognitive analyses of the OIBs. The next portion of the session consisted of a presentation by Dr. Lewis who provided a more detailed description of the bookmark procedure and specific information and instructions on threshold students, bookmark placement for Level 3, Level 2, and Level 4. New slides were added; among them, one with the specific “Bookmark Placement Instructions” and the additional slide and explication of “consistently” that had been discussed in the debriefing session the previous afternoon with attention was given to issue of "consistently" vs. 50% chance of answering correctly. It was not clear that the issue was totally cleared up for all participants (although in observations of subsequent grade/subject discussions and activities, any lack of clarity did not appear to affect participants' ratings.

After dismissal from the whole group meeting, participants returned to their grade/content rooms to review/complete the cognitive analysis of the OIBs. The additional 45 minutes allocated to the task appeared to be helpful to the groups. After a break, the groups then began the practice activity in which they placed a Level 3 (only) bookmark using a six-item orientation-round OIB and other resources. Results from the practice were shown to the groups, and a review of how to place the bookmark, discussion of the content of the item that contributed to the placements, and clarifications were offered.

In the grade/subject rooms, the facilitators then provided directions on how to set the operational bookmarks for all levels, beginning with Level 3, then Level 4, then Level 2. Panelists were instructed to place their bookmarks independently at their tables and to complete the first questionnaire upon closing out of the practice round before beginning their Round 1 operational bookmark placements. The Round 1 bookmark placements continued through a break for lunch at approximately noon.

Grades 6-8 Sessions, Day 2: Afternoon Activities (October 16, 2014)

The afternoon session of Day 2 began with most grade/content groups finishing up Round 1 ratings or, for groups that had finished prior to lunch, beginning a review of the Round 1 results and engaging in table discussions about where individual panelists had placed their bookmarks. In each of the groups, the item maps showed a column called "Location" which showed the scale locations of each item in the OIB. The way this information was treated across the groups appeared to vary, with some facilitators providing minimal attention to that column, whereas at least some of the facilitators provided information to panelists on how to use the information. In at least one group (Grade 8 ELA), the facilitator slowly and explicitly helped the panelists in that room to see how moving a bookmark several pages in an area where the items were close in their scale locations would minimally affect the resulting cut score, whereas moving a bookmark even one or two items in a location where the items differed more substantially in their locations would have a more pronounced effect on the cut score for that level.

As part of the Round 1 review, table groups were instructed in how to determine their median benchmark score and were asked to compare their table median to the whole group median and discuss rationale for placing bookmarks. Discussion focused on Level 3 bookmark first, and then Levels 2 and 4. After a thorough discussion of bookmark placements, results from the online panel were presented and discussed. The online panel results were grouped according to participant self-designated category: Teachers, Administrators, Other (parents, etc.). Medians and interquartile ranges were shown for each major group and the overall group. Participants discussed the online participant results, but were directed to use them as another point of reference only, not something that should distract their focus from the ALDs for setting their bookmarks. Participants then completed a readiness questionnaire before beginning their Round 2 bookmark placements. Round 1 discussion times varied among groups, but all groups took a short break between 3:00 and

3:35 p.m. All groups finished their Round 2 bookmark placements before 5:00 p.m.

At the end of the day, a debriefing session took place attended by contractor staff, Smarter Balanced staff, and the auditor. No major issues were raised. Dr. Lewis indicated that the addition of the "Bookmark Placement Instructions" had been beneficial and had helped participants clarify and focus their efforts. He also provided guidance for Day 3 by asking facilitators to present all of the information on the slides in the morning before going into deep discussion on any one slide. There was also a brief discussion about strategies for presenting the impact data and focusing participants on determining appropriate cut scores based on the ALDs and not letting them get too focused on the impact data and "chasing numbers."

Grades 6-8 Sessions, Day 3: Morning Activities (October 17, 2014)

The morning session of Day 3 for grades 6-8 began at 8:30 a.m. with facilitators convening their grade level groups. After going over logistics such as shuttle times to the airport and checkout time, the whole group began review of Round 2 data and impact data. All major technology issues seem to have been resolved as none were observed in any of the sessions. Normative and impact information provided were the same as the data provided to the grade 11 groups, with minor exceptions. The complete information provided to panelists consisted of the same six slides (described below) for their grade levels. For each group in the grade 6-8 band, a seventh slide which showed the impact data for Round 3 of the grade above. (The exception was grade 8, which saw the impact data from grade 11, Round 3.)

Slide 1: Histogram of individual bookmark placements. In many groups, there appeared to be good separation between levels and no "reversals" (e.g., instances where some group members placed their Level 3 bookmarks at locations that were lower than other panelists placed their Level 2 bookmarks). Discussions focused on the range of bookmark

placements within levels. In some groups there was overlap between levels, with the high end of one level overlapping with the low end of the next level. In these instances, the facilitator focused discussion on the overlapping points and having the panelists who placed their bookmarks in this area explain their rationale. In all rooms, the focus was brought back to the ALDs and what the ALDs said about what the threshold student should be able to do.

Slide 2: Table and overall group medians and ranges. In a couple of groups, there was a table that generally set their bookmarks lower that the rest of the group and one that set their bookmarks higher than the rest of the group. It seemed that in these cases, the groups had identified an outlier item that matched an ALD description, but was not a part of a group of similar level items.

Slide 3: Impact data for Round 2. Data used were student performance on Spring 2014 field test and showed the percentage of students who would achieve. Level 1, 2, 3 or 4 based on panelists' bookmark placements. In general, panelists did not seem surprised by the impact data. They raised questions about how this impact data reflect typical first-time standard setting versus later performance. They expected that performance would improve as students and teachers become more familiar with the standards and tests,

Slide 4: Histogram of impact data

Slide 5: Online panel (Teachers, Higher Ed, Administrator, Other, All) impact data levels for students achieving Levels 1&2 and Levels 3&4

Slide 6: Histogram of students at Level 3 disaggregated by various demographics (ALL, males, females, Hispanic, American Indian, Alaskan Native, Asian, African American, White/Caucasian, Hawaiian/Pacific Islander, Multi-ethnic, IEP, LEP/ELL, 504 Plan, Economically Disadvantaged)

Slide 7: Impact data for the grade above (e.g., Round 3 grade 11 for the grade 8 group, Round 2 grade 8 for the grade 7 group, and so on)--Levels 1&2 percentage and Levels 3&4 percentage

In most small groups, the information in all of the slides was presented and explained first. The full group discussion/analysis of the information took place after the presentation of the last slide. In some groups, the impact data raised concerns about potential uses of the test scores. For example, if too few students achieve Level 3 or 4 in some states, teachers may lose their jobs. Facilitators emphasized that the participants are making recommendations only and their recommendations are just one step in the standard setting process. Whenever such concerns were raised, facilitators brought the focus back to their task of placing bookmarks at appropriate points based on the ALDs. In one room, a participant made the distinction between thinking about where they want college and career ready students to be (what the ALDs indicate) and where their students are now (impact data). Other participants discussed motivational factors possibly suppressing student performance on the field test and recognition that in many field test states, the standards had not been fully implemented, so students may not have seen some of the material on the test. These discussions all helped participants contextualize the impact data showing fewer students at Level 3 and 4 than they would have liked to see.

After thorough discussion of Round 2 data and rationales for bookmark placement, participants completed a readiness survey and then commenced Round 3. (Participants were allowed to check out from the hotel during the interval between completing Round 3 and presentation of Round 3 results.) Facilitators stressed the confidential nature of the material presented and discussed during the meeting and reminded participants to leave their packets in the room. In one room, during the interval between submitting Round 3 bookmarks and receiving the final results, the facilitator led a discussion of "take-aways" from the session, what participants

might be asked by people in their home states, and how they might respond to difficult questions without divulging confidential information. Finally, each group was shown their group's Round 3 bookmark (median of the group) and impact data for the round then asked to complete a final questionnaire. By approximately 12:00, all sessions had adjourned.

Grades 3-5 Sessions, Day 1: Morning Activities (October 17, 2014)

Concurrently with the Grade 6-8 panels completing their Round 3 ratings, the standard setting session activities for Grades 3, 4, and 5 ELA and mathematics began. The activities for grades 3-5 followed the same agenda for table leader training on the prior evening (October 16, 2014) and for the daily activities as had been followed for grades 6-8 and 11. After an opening breakfast, grade 3-5 participants met as a full group for the orientation session in the ballroom. The whole group orientation session began with a welcome and overview presentation by Dr. Willhoft including an overview of the general purpose of the sessions, provision of some background and context, questions, solicitation of participants' concerns; the session continued with a presentation by Dr. Bunch, who described the agenda and goals for the workshop, along with brief attention to background on the common core state standards (CCSS) including claims and targets, the item and test development process, assessment item formats, the ALDs, the practice test, the bookmark procedure, the concept of an OIB, performance tasks, evaluations, "key shifts" embodied in the CCSS, and the four types of ALDs with an emphasis on the threshold descriptions. Dr. Bunch ended his presentation by noting the "ground rules" for the workshop and by providing information on the room assignments and facilitators for the mathematics and ELA groups; these included:

Grade 3 ELA - Craig Deville and Sheryl Grady (MI)

Grade 3 Math - Heather Farina and Rick Mercado (CTB)

Grade 4 ELA - Kelly Connelly (MI) and Sarah Hagge (CTB)

Grade 4 Math - Winnie Reid and Lisa Johnson (MI)

Grade 5 ELA - Joe McClintock and Ruth Hargis (MI)

Grade 5 Math - Judy Hickman and Jennifer Lord-Bessen (CTB)

In the breakout sessions, facilitators provided background information on the Smarter Balanced assessment framework, CAT testing and three types of ALDS: policy, range, and threshold. Next, participants were shown how to access reference materials on the computer and were asked to bring up the CCSS and DOK pages. Facilitators distributed copies of the threshold ALDs and discussed their purpose. Table groups spent 10-20 minutes discussing the ALDs and, after the small group discussions, table leads shared key points that arose in their small groups. Table groups approached the ALDs discussions differently: Some focused on understanding a target/standard as it is articulated across the levels (e.g. fractions at level 2, then 3 and 4); others honed in on one ALD level and worked to understand all aspects of that level. n G4, math participants commented that some standards (e.g. time) were addressed in ALD 2 and 4, but not ALD 3. Some participants commented that they would like to see examples of what is meant by some of the terminology (e.g. low-to-moderate text complexity); other participants asked about how the reading level of the texts was determined.

In general, between 10:00 and 10:30 a.m., all groups took a short break and then reconvened to engage in preparation for the short practice test. The practice test is publicly available on the Smarter Balanced web site as a way to familiarize potential test takers with the variety of items types that will appear on the operational assessments. The practice test again provided participants with an opportunity to view the variety of online supports available to students. After taking the practice test, participants discussed the items in small groups and then

participated in a large group discussion of their observations. During this subsequent discussion, participants commented that students would need to be trained on test taking strategies and become familiar with appropriate technology. Examples such as knowing how to scroll, ability to type, knowledge of the tools available to them (e.g. glossary) came up as things students would need to learn to do to perform well on the tests. After the group discussion and before adjourning for lunch at 12:30 p.m., participants took a brief readiness survey.

Grades 3-5 Sessions, Day 1: Afternoon Activities (October 17, 2014)

After lunch, participants reconvened in the large ballroom. The whole group activity in the afternoon began with Dr. Willhoft giving brief comments and informational updates. The session then included the same presentation by Dr. Lewis as was provided on the afternoon of the first day for the grades 6-8 and 11 panelists. That presentation, "Understanding What the Test Measures," provided an overview of the item map and the tasks that participants would be performing in the afternoon, and emphasized the importance of answering the two main questions about each item (i.e., what the item measures/sources of challenge, and what makes an item in the OIB more challenging than the preceding items.) Dr. Lewis indicated that, over the next couple of days, participants would use this understanding to place Level 2, 3, and 4 bookmarks according to the ALD information; he also emphasized that participants should focus on the content of each item in the OIB with the goal of understanding what they know about what a student can do if the student gets the item correct. As part of his presentation, Dr. Lewis walked through two performance items to illustrate the various pieces of information provided about each item (e.g. DOK, target standard, rubric, student exemplars) and to emphasize that each performance item stimulus would appear for each score point. For example, an item worth two points would appear in the OIB twice: once for score point 1 and once for score point 2.

At approximately 2:00 p.m., panelists were dismissed to work in their grade/subject area groups where they began the process of examining the OIB. At the beginning of the small group session, participants asked a number of questions. Many were answered immediately, such as general housekeeping items and basic questions about Smarter Balanced, composition of ALD-development groups, and computer adaptive testing. More detailed questions were referred to Smarter Balanced staff (e.g. weighting of components of test). In some cases, participants brought up questions about the various uses of the tests in their states (e.g. teacher evaluations, promotion decisions). Facilitators focused participants back on their task of understanding the OIB well and placing appropriate bookmarks for the thresholds based on the ALDs.

During the afternoon session, a Smarter Balanced staff person (Nikki Elliott-Schuman, Smarter Balanced ELA content lead) circulated to all of the ELA small group sessions to address questions collected from participants in the morning session. (The Smarter Balance mathematics content lead, Dr. Shelbi Cole was also present and supported the mathematics panels.) Getting questions answered seemed to reassure participants who have come to this process from a wide variety of background experiences with the CCSS and Smarter Balanced assessments. In response to participant questions, Ms. Elliott-Schuman also shared information about enhancements that are being explored/developed such as a training test to help students understand and use the various accommodations available on the tests.

Facilitators then walked participants through the process of reviewing the OIB, had them log in, bring up the booklet and together review one or two questions. Facilitators picked example items to illustrate various aspects of items that participants would encounter and showed them how to enter comments to refer back to over the next few days as they work on placing bookmarks. It was strongly emphasized that the comments were for participants' eyes only; facilitators and other staff could not read them. In the ELA sessions, facilitators reviewed the presentation of a

performance assessment item, which appears six times in the OIB, twice for the conventions dimension (points 1, 2) and four times for the organization/use-of-evidence dimension (points 1, 2, 3, 4). For each score point, participants could access a rubric and student exemplars with annotations explaining why the response earned the specific score. Facilitators addressed all questions about the afternoon activity and made sure that participants were navigating smoothly within the OIB. Participants used the rest of the afternoon—until approximately 5:00 p.m.—to work through all of the items and make comments on them. Although the session officially ended at 5:00, participants in some rooms were allowed to review items until 5:30. In several other rooms, participants were not able to complete their initial review of the OIB. Some groups ended at 5:00 and none went beyond 5:30 p.m.

At 5:30 p.m., the end-of-day debriefing session took place, attended by contractors, Smarter Balanced staff, and the auditor. The time required to complete the initial review of the OIB was again raised as an issue; it was reported that only the grade 4 and grade 5 math groups had come close to finishing their initial review. In the previous sessions, it had worked well to allow time in the morning of Day 2 to continue review of the OIB; the same procedure was suggested for this group. It was decided that, after the morning large group session, each grade-level room would continue OIB review for 30-60 minutes before moving on to their next task.

Grades 3-5 Sessions, Day 2: Morning Activities (October 18, 2014)

The second day of the Grade 3-5 sessions began with panelists meeting in their grade/content groups to finish the cognitive analysis of their OIBs. The schedule adjustment adding this time appeared to be appropriate, as many panelists used the time to complete the work they began from the previous day. After completing the analysis, all groups experienced the training test and practice test activities, including discussions of why they placed their practice

Level 3 bookmarks as they did and other content focused discussions. Placement of Round 1 bookmarks began before lunch on Day 2, with some groups finishing the Round 1 bookmark placements before lunch and others taking a break for lunch and finishing up the Round 1 placements immediately after lunch. All groups who completed Round 1 bookmarking also completed a questionnaire and viewed information on Level 3 bookmark placements generated by the on-line panel for their grade and subject. In general, it appeared that connectivity issues were cleared up and few tech disruptions occurred.

Grades 3-5 Sessions, Day 2: Afternoon Activities (October 18, 2014)

The afternoon session of Day 2 began with most grade/content groups finishing up Round 1 ratings or, for groups that had finished prior to lunch, beginning a review of the Round 1 results and engaging in table discussions about where individual panelists had placed their bookmarks. In one room, there was somewhat of a lengthy delay in getting the on-line data to review. The facilitator sought help from the IT support staff and encouraged panelists to continue discussions of their Round 1 bookmark placements during the interval. In addition to the on-line information, the group median and range of bookmark placements were also provided in tabular and graphic formats. The rest of afternoon was spent setting Round 2 bookmarks, with panelists again instructed to set their bookmarks for all three levels in the appropriate sequence, submit their judgments, then fill out the end-or-round questionnaire when they were finished. Many panelists appeared to complete the Round 2 bookmark tasks early, with all finishing the tasks by 4:00 p.m. They were instructed that the activities for Day 3 would begin at 8:30 a.m., and asked to report to the individual grade/subject breakout rooms for the start of Day 3.

As had been done previously, Day 2 for grades 3-5 ended with a debriefing session involving contractor staff, Smarter Balanced representatives, and the auditor. Because all

participants had finished their Round 2 ratings somewhat early, the debriefing for grade 3-5 was convened a bit earlier in the afternoon on Day 2 at approximately 4:15 p.m. The group was informed that one participant had a family emergency and needed to leave early to return home; some attention was also given to logistics, such as moving the lunch break and shuttle service for Day 3 a bit earlier. Dr. Bunch queried the group for any common concerns; no major issues requiring attention or modification were identified. Because it was anticipated that the groups would have ample working time on Day 3, Dr. Bunch directed the facilitators to take time reviewing bookmark placements, to encourage thoughtful group review and discussion of any overlaps or "reversals" in bookmark placements (e.g., situations in which a group member's Level 3 bookmark was placed at an earlier location than another member's Level 2 bookmark).

Grades 3-5 Sessions, Day 3: Morning Activities (October 19, 2014)

The activities of grades 3-5 on Day 3 consisted of a review of Round 2 results, a questionnaire, placement of Round 3 bookmarks, presentation of impact data and a final questionnaire. Two groups were observed in depth on this morning, and the observations provided the basis for our conclusion that facilitators in different rooms had differing degrees of experience and comfort leading standard setting workshops. In one of the groups observed, the facilitation was masterful. The facilitators explained all information clearly; they responded accurately and appropriately to all participants' questions; they anticipated concerns; and made highly appropriate and useful time out of intervals when they were awaiting data analysis results. When participants had already completed their Round 3 judgments and were waiting for their data to be analyzed and returned, the analysis took a fairly long time—perhaps as much as 30 minutes—however, the facilitators made very effective use of the wait time, thanking participants for their sacrifice and effort, for giving time over the weekend, and reminded participants to share

back home about the ALS process, but no to provide information about people, items, or results. The facilitators also provided information on logistics (shuttles, lunch, etc.). When the data analysis was complete, results for final review and evaluation were projected, including the room's median Round 3 Level 2, Level 3, and Level 4 bookmarks, and impact data for all four levels (table and bar chart). There was some discussion of the results, but participants generally seemed to feel that the results were appropriate. After completing the final questionnaire, participants returned their paper materials, and were dismissed for lunch and return trips home.

In another of the groups observed on this morning, the facilitation was less skilled and seemed more labored. The session began with presentation of a histogram showing Round 2 bookmark placements for the group, along with the group's interquartile range and median. A second slide was shown that provided the Round 2 median cut scores for each achievement level for the five table groups and for the total grade/content area group. The lead facilitator in the room seemed to have difficulty understanding and explaining the median values shown on the slide, asking out loud if higher medians or lower medians reflected higher or lower expectations. The third slide shown provided impact data (percent of students that would be categorized at each achievement level) for the total room based on field test performance. The lead facilitator read the interpretation of the results shown on the slide; the second facilitator elaborated on the interpretation and aided the group in understanding how they should use the impact data. A panelist asked how the impact percentages would change if the cut score for a level were increased by one raw score point correct. The facilitator indicated that it was not possible to tell based on this data. Discussion was then prompted regarding reactions to the impact data, with participants in general reacting to a comparatively larger percentage of students being classified as Level 1 and a comparatively smaller percentage of students being classified as Level 2. The facilitator then reposted the histogram (slide 1) and related the spread of bookmark locations to those

percentages--a comparison that was essentially irrelevant and potentially misleading, conflating two results (i.e., range of panelists' judgments and categorical impacts). Again, the second facilitator in the room attempted to provide some additional information that may have helped panelists understand the results. One panelist contributed an interpretation that was accurate and provided a clear way for panelists to think about how they would move their bookmarks to affect the percentage of students in a category. However, the panelist also suggested that the task of the group was to obtain a "standard normal distribution" of results and that the current impact did not reflect that—a suggestion that the facilitators did not correct.

The next slide shown provided the medians and interquartile range from the on-line panel for the Level 3 cut, showing the impact (percentage of students that would be classified as Level 3 or above). Another slide containing impact for subgroups (percent that would be classified as Level 3 or above) was also shown. Finally, the group was shown the impact (percent in Levels 1 and 2, and percent in Levels 3 and 4) based on the upper grade (Grade 6) round 3 final recommendations. The facilitator asked the group to "gasp in unison" at the discrepancy between the results for the two grades (i.e., grade 5 and 6). Impact data based on the lower grade was not provided. The facilitator then explained that the next step in the process was that a vertical articulation panel would review and potentially smooth out differences in impact across grades and the panelists were encouraged to focus on content. A panelist reinforced that the task of the group was to "focus on the ALDs and how those are translated into their expectations instead of trying to match another grade's results."

Panelists were then instructed to open their OIBs and review their bookmark placements as shown on the first slide (histogram) and to share their rationales for why they placed their bookmarks at those locations, particularly for the Level 2 cut. As the discussion started to drift, a panelist reminded the group that the focus should not be on what they think the percentages should

be, so much as what the ALDs indicate. The panelists then discussed the other levels in like manner. For the level 4 discussion, a panelist defended her placement of a bookmark distinguishing level 3 and level 4 by saying that she thought the upper level 3 students would be able to get the item correct, but was asked by the facilitator, "would at least 50 percent of them get it correct?" Although some of the strategies and understandings revealed in the discussion seemed incorrect, overall the group appeared to have a good grasp of how to place their bookmarks. The group began a break at 10:35 a.m. then returned to complete a questionnaire, setting of Round 3 bookmarks, reviewing Round 3 results; they were then thanked for their participation and dismissed.

IV. Vertical Articulation

A vertical articulation session (called "Cross Grade Review") was held on Monday, October 20, 2014. The session began at 8:30 a.m. Dr. Bunch welcomed the group and introduced Deb Sigman and Joseph Martineau (Smarter Balanced Executive Committee Co-Chairs) who both also welcomed the group, thanked them for their work, and gave brief remarks on the purpose of the session—ensuring a coherent system across the grades. Dr. Bunch also introduced the group facilitators for the ELA and mathematics panels (himself and Dr. Lewis) and he introduced Dr. Willhoft who also gave welcoming remarks an provided an overview of the next steps in the ALS process. The articulation panelists were distributed as follows:

ELA Cross Grade Panel (32 total panelists)	Mathematics Cross Grade Panel (32 total panelists)
Grade 11 - 16 panelists	Grade 11 - 16 panelists
Grade 8 - 2 panelists	Grade 8 - 2 panelists
Grade 7 - 3 panelists	Grade 7 - 3 panelists
Grade 6 - 2 panelists	Grade 6 - 3 panelists
Grade 5 - 3 panelists	Grade 5 - 3 panelists
Grade 4 - 3 panelists	Grade 4 - 3 panelists
Grade 3 - 3 panelists	Grade 3 - 2 panelists

Dr. Bunch then presented a series of slides and information that provided participants with an historical overview of each phase of the process performed to date and those to come, beginning with the on-line panel, the in-person workshop, a TAC review, and culminating with review and approval by the chief state school offices on November 6, 2014, and production of a technical

report. The next slide showed the progression of review from individual, to table, to grade/content, to cross-grade. The third slide provided answers to the question "Why are you here?" and he described the efforts made to ensure broad and qualified representation. The fourth slide outlined the participants' tasks: 1) examining cross-grade bookmarks; 2) establishing cross-grade coherence and reasonableness; and 3) reviewing impact data and scaled scores. The next slides provided elaboration on the definition, requirements, and typical patterns of cross-grade articulation, along with an example of an "unexpected" pattern of cross-grade standards.

The next slide introduced the rationale for articulating impact data and vertically-scaled scores across grades, as opposed to bookmarks. Graphs showing the mean scaled scores on the vertical scale across grades that resulted from the field test data were then shown for ELA and mathematics. The presentation then turned to "The Tools We Will Use" which explained the presence of variability within grade/content panels and introduced box and whisker plots to represent that variability. Tables showing hypothetical data of grade level scaled scores, impact data, and the relationship to bookmark placements were presented to familiarize participants with the data they would be using for their cross-grade articulation task. Dr. Bunch reviewed materials that panelists would have for reference, including all of the grade/content OIBs and their grade/content ALDs.

The specific steps that participants should follow were then presented. The steps included review of current bookmark placements, consideration of moving bookmarks at grade levels, and voting procedures for group acceptance of a proposed bookmark change. Procedures included the requirement for making and recording of formal motions to suggest a change, a second for the motion, and a 2/3 majority required for approval of a suggested change with a recorded tally of "yeas", "nays" and abstentions. It was not clear if a vote would be required if a panel chose to ratify the original bookmark location. (This auditor mentioned this issue to Mr. Alpert, who

indicated that it would be addressed in the breakout groups.) Dr. Willhoft reviewed the requirements for who would be eligible to vote. A slide was then shown that illustrated the voting log that groups would be asked to complete. There was some lack of clarity regarding who was eligible to make a motion; Dr. Bunch indicated participants he would "prefer" for motions to be made by someone from the grade level under consideration, although he also indicated that a motion could be made by anyone eligible to do so fro the total group. It was also not clear if a motion was required to keep a bookmark location as originally recommended by a grade/content panel.

A short time of questions and answers followed at the end of Dr. Bunch's presentation. Dr. Willhoft answered largely policy and procedural questions from participants related to how the chief state school officers would make their decisions on final cut score adoptions and the information that they would be provided to aid in that decision. A brief logistics update was provided, covering time frames for the agenda, breaks, and assignments for the breakout rooms. The whole group was then dismissed at approximately 9:45 a.m. to begin work in their breakout rooms, one for ELA and one for mathematics.

In the breakout groups, panelists introduced themselves and gave brief descriptions of their roles. The ELA group was facilitated by Dr. Bunch and Dr. Craig Deville from MI; the mathematics group was facilitated by Dr. Lewis and Dr. Rick Mercado of CTB/McGraw-Hill. In general, it appeared that the optimal way in which facilitators led their groups was for one of the facilitators to lead the process, while the other facilitator processed input into the computer and ensured that the relevant materials was displayed.

Participants in each group were then shown the cross grade distributions of impact data; the graphs were shown with points along the grades indicated by a point showing the panel median with accompanying measure of variability (box plots with interquartile ranges and whiskers

extending from P10 to P90). The data and graphics were presented in a dynamic fashion, so the effect of any suggested changes could immediately be seen in terms of impact and scaled scores.

Discussion of the results followed. In both content areas, the groups began their work by examining the articulation between the upper-most grades (i.e., grade 11 and grade 8), then continued their review working downward across the grades. The groups appeared to understand the nature of the data presented to them and the task they were empaneled to perform. However it also appeared that panelists believed that they needed to smooth out all bumps and dips in the impact data. Groups began the work of considering changes and broke for lunch at approximately noon, returning to complete their work at approximately 1:00 p.m. The groups began by working on articulation of Level 3 across the grade levels. After completing that work, the proceeded to consider articulation of Level 4, followed by Level 2.

Overall, the facilitation in the breakout rooms seemed to be highly effective. Facilitators answered panelists' questions clearly and kept the process on track. In one of the rooms, at the beginning of the breakout session, at least some panelists commented that they felt somewhat rushed in their work. To a small degree, in both rooms there was some disagreement among panelists about the appropriate referent for their decisions—the ALDs, the state of implementation of the common core in their states, and where student performance is currently and the impact on students of the decisions. However, for the most part, panelists considered adjusting impact, grounding their judgments in the content of the OIBs and the ALDs. The discussions were largely content focused; all participants appeared to be comfortable contributing to them; and no personal agendas were evident. For the most part, the content-based adjustments made by the panels fell within the P10 to P90 range from the original panels. In rare instances, the panels exceeded those values. Whenever a change was voted upon, the groups generally worked to endorse a change that was broadly acceptable; it was typical for votes to be unanimous within a group, or nearly so. All groups

finished their work by 5:00 p.m. An evaluation questionnaire was not administered to the vertical articulation panelists. They were thanked for their contributions and dismissed.

V. Summary and Recommendations

Based on observations of the procedures and processes used to obtain recommended performance standards, it is my opinion that the standard setting activities implemented for the Smarter Balanced summative assessment standard setting were, overall, conducted in a manner consistent with sound psychometric practices.

Few issues arose during the in-person standard setting sessions or the vertical articulation session; issues that arose were comparatively minor and, in my opinion, would not have substantially affected the validity or reproducibility of the results. Although there are several aspects of the process that should be considered for revision and incorporation for future standard setting procedures, none of them could be considered to be a fatal flaw in the process. In the following subsections, some specific strengths and recommendations are presented.

Strengths

There were a number of strengths observed during the standard setting activities for the Smarter Balanced assessments. A thoughtful standard setting plan was developed and reviewed by the Smarter Balanced Technical Advisory Committee and formally approved by Smarter Balanced governing states. Overall, there was a well-organized and faithful implementation of that plan by the contractors responsible for conducting the standard setting workshops. The contractors provided adequate resources and personnel to ensure that the standard setting was conducted professionally and paced appropriately.

For the in-person sessions, it appeared that all panelists had strong qualifications for participation; they appeared to be highly qualified and they seriously, conscientiously, and

enthusiastically engaged in their tasks; they generally understood how to complete their tasks and they engaged in relevant, content-focused discussions. All participants appeared to work diligently during the sessions; no issues regarding domination of discussion in groups/tables were apparent, and no participants appeared to exert personal agendas.

With the exception of some of the feedback regardin external data sources (e.g., PISA, NAEP, ACT), participants appeared to understand the nature of the feedback provided to them (i.e., normative and impact information). In discussions they appeared to reference and consider the information appropriately. Overall, participants identified as table leaders appeared to function well in both the initial standard setting sessions and the subsequent vertical articulation session.

For the vertical articulation process, the panelists were similarly representative, engaged, deliberate, and thoughtful in their work. The orientation session appeared to be carefully planned and well delivered; the breakout sessions were also facilitated exceptionally well. At the beginning of the breakout sessions, some panelists commented that they felt somewhat rushed in their work. For the most part, all the discussions were content focused; all participants appeared to be comfortable contributing to them; and no personal agendas were evident. There was some disagreement among panelists about the appropriate referent for their decisions—the ALDs, the state of implementation of the common core in their states, and where student performance is currently and the impact on students of the decisions.

In addition, the following specific strengths were observed:

1) All procedures were pilot tested and adjusted in response to experience from the pilot;

2) Improvements in the operational standard setting based on lessons learned from the pilot testing were incorporated (e.g. hard copies of ALDs provided, doing the practice test, more

focus on making participants aware of item aspects, and so on);

3) The process was marked by substantial transparency (e.g., parking lot answers to any questions raised by the groups, whether directly related to standard setting or not);

4) The technology generally worked well, including computers, headphones, projectors, and software used for the sessions. Comparatively minor technology issues were quickly and effectively addressed;

5) The groups were divided into tables to facilitate more ready and deeper discussions;

6) A cognitive analysis of the full OIBs was included; it allowed panelists to become deeply familiar with the test content on which they would be making their judgments;

7) A practice activity was included for panelists to gain greater familiarity and experience with the bookmarking procedure;

8) Adequate time was allocated to allow the procedures to be completed at an appropriate pace. The schedule adjustment to allow additional time for all groups on Days 2 and 3 was a helpful adjustment, allowing participants to complete their cognitive analysis of items in the OIB. Had this adjustment not been made, many participants may not have become familiar with the items at the end of the OIB--which would represent the most challenging content and, likely, content tapping some targets/standards not represented previously in the OIB;

9) Nightly debriefing sessions occurred to gather feedback from facilitators on the day's activities and to plan for appropriate adjustments;

10) Appropriate evaluations ("Reflection Questionnaires") were administered at relevant

junctures in each in-person workshop;

11) The TAC-reviewed and state-approved plan was essentially followed faithfully; only minor deviations from the plan as written were observed. Some of the deviations were unintended (e.g., one group did the practice test out of order) and some were intended (e.g., inclusion of new slides on bookmark task); none of the deviations appeared to be consequential in terms of affecting rigor/integrity of process, results;

12) The orientation and training covered some basic test development information and the role of the ALDs was emphasized throughout the process;

13) All procedures went well on Day 1 of the first session; all procedures improved over the course of the three workshops as minor improvements made things run more smoothly.

14) The logistics were handled exceptionally well. The hotel accommodations, staff, food service, and other aspects of the venue were without any issues. The contractor staff responsible for logistics also helped the meeting proceed without a hitch: panelists always knew where to go, when, had relevant materials, and were assisted with all arrangements, such as transportation and special needs (e.g., a larger monitor was quickly obtained for panelist with visual impairment);

15) There appeared to be capable and responsive technical and practical support that provided accurate and timely data analysis and feedback for panelists;

16) The materials, forms, and documents appeared to be well-designed and easy for participants to use; and

17) There was appropriate concern for and attention to confidentiality and security of

materials and results, including the collection of all paper materials at the conclusion of the panels' deliberations.

Limitations/Considerations for the Future

A few areas may warrant attention either as information for policy makers, or as information that Smarter Balanced or contractors may wish to consider when planning or implementing future standard setting activities. Some observations and recommendations include:

1) The computer interface seemed somewhat challenging for participants to use, often requiring clicking back and forth among screens or documents or several click sequences to get to desired materials. In the current standard setting, participants appeared to adapt to the interface, but greater intuitive look and ease of use would seem desirable.

2) The facilities for the present standard setting were marked by small rooms, rectangular tables, and equipment on the table (monitors) that made it difficult for groups to work independently and difficult for participants at a table to converse among themselves. Again, participants adapted in various ways (e.g., by relocating their monitors, moving their chairs into circles away from their tables, etc.), but larger rooms, lower footprint monitors (laptops?), or other feasible configuration changes would enhance independence among groups and discussions within groups.

3) There appeared to be recurring, intermittent connectivity issues. Although the participants' work was not substantially impeded and technical support staff always quickly and efficiently addressed the issues, it would seem that more stable/reliable connectivity should be investigated for future studies.

4) The auditors were not able to see all activities for all rooms. It is recommended that at least two auditors be assigned the task of performing observations when future standard setting workshops involve multiple groups working concurrently.

5) There appeared to be fairly wide variability in how ancillary information (e.g., NAEP, PISA, ACT, on-line data) was presented and interpreted in the groups. Information was presented and described at different times, in different ways by different facilitators. Some facilitators indicated that participants should "feel free to use, incorporate, review, ignore [the information] as you see fit;" others facilitated thoughtful discussions on the data. For the future, it would be desirable for workshop planners to careful consider precisely *how* participants should consider and incorporate any such data sources and more uniform scripting and implementation of this aspect of the process would be desirable.

6) There appeared to be substantial variability in facilitation of the breakout groups: some facilitators appeared to be highly experienced, thoroughly understood the process and concepts, took charge, elaborated on script to help participants fully understand/perform their tasks, led very effective discussions and so on; others seemed to be much less experienced, lacked a level of comfort in conducting the sessions, were less familiar with the methodology, materials, agenda, software, output, stayed very close to scripts or read scripts verbatim, and lacked skill in facilitating deep discussions when results were presented. This concern was exacerbated when rooms were staffed by two less experienced/comfortable facilitators. For the future, it is recommended that attention be paid to ensuring at least one facilitator in each room should be highly skilled and comfortable with the activities; in no case should two less-seasoned facilitators be assigned to the same group.

7) There was some variability in how table leaders functioned. Some table leaders facilitated more effective discussions than others; some groups functioned more or less democratically, and so on. For the future, it may be desirable to provide greater training to table leaders and for facilitators to have a debriefing session each day with table leaders to reinforce table leader roles, identify successful practices and to recommend strategies for addressing concerns raised by the table leaders.

8) In the cross-grade articulation process, it may have been desirable to also explain to panelists that mild to moderate fluctuations across the grades could be acceptable. In addition, some panelists commented that the data presented on some of the slides was too small to read. In the future, it would seem desirable to increase the font size for the presentations, or, if feasible, another strategy might be to "push" non-dynamic versions of the screens to participants' monitors such as is done in a webinar format.

9) In the cross-grade articulation panel, it would seem desirable to administer a final evaluation questionnaire. Such a data collection would have the potential to serve as additional validity evidence for the process.

Conclusions

Overall, I can provide a positive evaluation of the standard setting activities for the Smarter Balanced summative assessments. On the one hand, there were a few aspects of the process that represented concerns that weakened the confidence that could be gained from the process. On the other hand, there were a number of strengths apparent in the standard setting procedures. Importantly, the plan for setting the performance standards was developed in conjunction with the advice of the Smarter Balanced Technical Advisory Committee. The procedures were

implemented with good fidelity to that plan. Further, it is my opinion that the procedures and processes used to derive recommended achievement levels followed sound, best practices of the psychometric profession. All participants--educators, policy makers, public representatives, and contractor staff--appeared to take a serious and conscientious approach to the tasks. The inclusion of a vertical articulation activity provided an important "check" on the coherence and reasonableness of the panelists' recommendations.

It is my conclusion that the standard setting activities described in this report were designed and conducted appropriately so as to yield defensible performance standards grounded in the knowledge, skills, and expectations represented by the ALDs. Only minor issues arose during the standard setting and vertical articulation processes; in my judgment, none of them present a major barrier to the integrity of the results. Because the procedures and processes used to derive recommended performance standards appeared to follow sound psychometric practices. Unless analyses of the panelists' evaluations were to indicate otherwise (I did not review these prior to submitting this report), I conclude that the panelists' cut score recommendations should be considered to be valid and reliable estimates of appropriate, content-referenced cut scores to define the four performance categories on the Smarter Balanced summative assessment. I believe that policy makers can have confidence that the recommendations from the standard setting panelists are based on sound procedures, producing trustworthy and defensible results.

VI. References

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (Ed.) (2012). *Setting performance standards: Foundations, methods, and innovations.* New York: Taylor and Francis.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: SAGE.

Cizek, G. J., & Bunch, M. B., & Koons, H. H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Lewis, D. M., Mitzel, H. C., Mercado, R., Schulz, E. M. (2012) The Bookmark standard setting procedure. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (pp. 225-254). New York: Taylor and Francis.