Observation and Report on

North Carolina End-of Grade (3-8) , End-of-Course (NC Math 1, NC Math 3), and NCEXTEND1 (3-8) and NC Math 1 Alternate Assessment Mathematics Standard Setting Workshops

July 8-11, 2019

Marriott Crabtree Valley
Raleigh, NC

Prepared by:

Gregory J. Cizek, PhD

July 17, 2019

Executive Summary

The North Carolina Department of Public Instruction (NCDPI) contracted with Data Recognition Corporation (DRC) to engage in a process for deriving performance standards (i.e. "academic achievement levels," or "cut scores") for a series of North Carolina student achievement tests in mathematics. Three performance standards were identified to define four levels (*Level 5, Level 4, Level 3* and *Level 2 and Below*) for the North Carolina general population End-of-Grade (3-8) and End-of-Course (NC Math 1, NC Math 3) assessments;, and two performance standards were identified to define three performance levels (*Level 4, Level 3*, and *Level 2*) for the North Carolina alternate assessment population (End-of-Grade 3-8, NC Math 1). Because the North Carolina State Board of Education is the entity with the authority to actually set the performance standards on NC state assessments, it is important to note that the process resulted in recommendations for performance standards that would subsequently be reviewed and potentially approved by the relevant state personnel.

Large, diverse, and representative groups of qualified participants—namely, North Carolina classroom educators—were empaneled to perform the standard setting procedures on July 8-11. 2019 at a large conference hotel in Raleigh, North Carolina. A total of 58 educators participated in the general assessment standard setting activities; 40 educators participated in the standard setting for the alternate assessment.

The sessions for each assessment program (i.e., general assessment, alternate assessment) were implemented using different standard-setting methods. The Bookmark procedure (Lewis, Mitzel, Mercado, & Schulz, 2012) was used for the general assessments; the Yes/No Angoff procedure (Plake & Cizek, 2012) was used for the alternate assessments.

Qualified educators from North Carolina were trained in the methods and led through the standard setting procedures by content and process specialists. The participants' judgments were solicited in two ways: they first generated exclusively content-based judgments and cut scores across three rounds of judgments in Phase I of the standard setting workshop; they next made adjustments to the system of recommended cut scores in vertical articulation sessions in Phase II of the workshop

The author of this report was contracted by NCDPI to provide an independent, external observation of the standard setting sessions and to submit a report of observations and findings. The author has expertise and extensive experience in the area of setting performance standards (see, e.g., Cizek, 2001, 2012; Cizek & Bunch, 2007). The author also serves as an NCDPI technical advisor; this allowed the author to review and become familiar with the standard setting plan as NCDPI desired it to be implemented and to evaluate any deviations from that plan, if they occurred.

Overall, the workshop produced cut score recommendations that can be considered to be valid and reliable estimates of appropriate performance standards for the relevant assessments. Unless the panelists' evaluations indicate otherwise, policy makers should have confidence that the recommendations from the standard setting activity are based on sound procedures, producing credible, defensible, and educationally useful results.

The remainder of this report provides a description of the standard setting activities, some recommendations, and a summary evaluation. The report is organized into four sections: 1) Executive Summary 2) Standard Setting Workshop Observations; 3) Summary and Recommendations; and 4) References.

II. Standard Setting Workshop Observations

Day 1: Morning Activities

The morning sessions on Monday, July 8, 2019 began at approximately 8:30am, with a welcome and orientation to the standard-setting task was provided. Before the formal welcome and opening remarks, DPI staff ensured that all required forms had been signed on check-in by all participants; a few participants who had not signed their forms were asked to do so before the session began.

In a whole-group setting, Dr. Tammy Howard (Director of Accountability Services) welcomed all participants, thanked them in advance for their willingness to serve, and provided with a general orientation of the upcoming standard setting activities. Dr. Howard introduced other DPI personnel, including, Ms. Kristen Maxey-Moore (Section Chief, Test Development), and DPI content representatives and psychometricians. She also informed participants that NCDPI had arranged for the standard setting workshop to be facilitated and supported by an external contractor, Data Recognition Corporation (DRC). Dr. Howard closed her opening remarks by reminding participants to focus on the important purpose of the meeting: recommending academic achievement levels to the North Carolina State Board of Education that help show state leaders, policy makers, educators, parents, and students how North Carolina students are performing with respect to the knowledge and skills covered in the state's curriculum.

Ms. Maxey-Moore also welcomed participants and provided the whole group with an overview of the test development timeline and process used in North Carolina. The 24-step process was illustrated using a PowerPoint presentation, and each step was briefly explained. Ms. Maxey-Moore described the three aspects of good assessment that drive the process: the tests developed in North Carolina must be valid. reliable, fair and meet all mandated constraints for timeline, budget, administration time, reading level, student development level and other factors.

Ms. Maxey-Moore then focused on one step in the process most relevant for the meeting: standard setting. The steps involved in the standard setting process were described, and the specific tasks the group would be performing were highlighted, beginning with an overview of the approved general assessment policy descriptions, developed by stakeholders from across the state. Four descriptions for the general assessments, one for each level, were described. The current labels for the general assessment levels are *Level 2 and Below, Level 3, Level 4*, and *Level 5*. The three current levels for the alternate assessment are *Level 2, Level 3* and *Level 4*. Ms. Maxey-Moore then differentiated between the intended student population for the general assessments and the NCEXTEND1 assessment population. Ms. Maxey-Moore also highlighted the performance levels related to academic preparedness of college and careers.

Ms. Maxey-Moore ended her portion of the introduction by covering some housekeeping details such as reimbursements, meals, honoraria, continuing education credits, access to personal cell phones, and other confidentiality and security-related issues.

After a brief pause, DRC mathematics content specialist Dr. Scott Woelber provided information relevant to the group's first task: creating grade-level Range achievement level descriptions (Range ALDs) for the general and NCEXTEND1 assessments. Dr. Woelber described the different kinds of ALDs (e.g., Policy ALDs, Range ALDs, Threshold ALDs, and Reporting ALDs) and their use. As a beginning activity, Dr. Woelber asked the group to review some of the Range ALDs (previously created by DRC and reviewed by NCDPI) and, with a colleague at their tables, to discuss differences in what they saw in those Range ALDs. It was noted that these previously-created Range ALDs were only provided as suggestions as to style, format, specificity, etc., and that panelists should not only create Range ALDs where none were provided, but should also feel free to revise the previously-created Range ALDs.

His presentation then covered three main areas:

1) definitions for key terms that would guide the group's efforts;

2) a review of the specific purpose of the meeting on the first day (i.e., developing Range ALDs) and three aspects that affect and can be reflected in the Range ALDs (the content covered by the test items, contextual factors in the items, and the cognitive complexity of the items), along with examples to illustrate those aspects; and

3) description of the specific task for the day—creating Range ALDs for all grades for each of the covered assessments—and logistics for doing so (e.g., computer login folder access, table clusters, group recording of their work, articulation across grades, and so on.

At the conclusion of this presentation, the group was dismissed for a morning coffee break. Following the break, participants were instructed to reconvene in two smaller groups (one for general assessment Range ALD writing, one for Extend1 Range ALD writing) in separate break-out rooms.

At approximately 10:00am, the two subgroups were formed for the purpose of creating specific Range ALDs. One subgroup comprised panelists who would be developing Range ALDs for the general mathematics assessments; this group was facilitated by DRC mathematics content specialist, Eric Jensen. The other subgroup comprised panelists who would be developing Range ALDs for the Extend1 assessments; that group was facilitated by Dr. Woelber. Participants were seated at round tables designed to accommodate approximately 6-8 persons at each table, and 6 laptops were provided at each table. The laptops were pre-loaded with materials needed by the participants for the Range ALD creation task.

A total of 59 participants were present for the general assessment Range ALD development; 40 participants were in attendance for the Extend1 Range ALD development.

In the NCEXTEND1 group, there were approximately six more participants in attendance than there were available laptop computers. This situation occurred because, contrary to most standard setting workshops, 100% of the invited NC educators and invited alternate NC educators (i.e., participants who were asked to attend in case a regular participant became ill, had a family emergency, etc.) arrived for the standard setting workshop. The situation was handled smoothly by workshop facilitators and technical support personnel, who reassigned some participants to different tables and encouraged participants to share laptops as needed, although in some situations the resulting table group sizes (the largest was n=11) seemed somewhat large for the kind of close collaboration and discussions intended.

After a welcome by the two subgroup facilitators, and brief, within-group introductions, participants were reminded of the Range ALD-creation task they were to perform; they reviewed the "starter" Range ALDs developed by DRC content specialists; and they reminded the groups to record their ideas for the Range ALDs for each achievement level on the pre-loaded templates provided on the laptops. The groups performed their work until a lunch break which was announced at Noon and scheduled for 45 minutes.

Over the lunch break, DRC and NCDPI staff met to discuss the progress of each group and how table leaders would be identified for the activities on Day 2. It was decided that table leaders would be identified differently for each group (self-identified by table members for the general assessment; identified by NCDPI/DRC staff for the alternate assessment). It was also decided that DRC would provide a brief training for participants identified as table leaders at the beginning of the lunch break on Day 2. Overall, it was concluded that the Day 1 morning activities had proceeded smoothly, with no major issues requiring adjustment. One minor issue arose involving two participants assigned to the Extend1 Range ALD reviews: the two participants indicated that

they were not interested in working on the alternate assessment area; these participants were reassigned to the general assessment group.

Day 1: Afternoon Activities

Two concurrent activities occurred on the afternoon of Day 1. First, members of each group continued their work drafting Range ALDs for their grade levels. In one subgroup, some participants who had been working within their individual grade levels were asked to take some time to confer with participants working on adjacent grade levels in pre-identified grade bands (grades 3-4, 5-7, and 8-NC Math 3). To accomplish the within grade band consultation, for example, selected grade 6 participants might spend some time reviewing and providing input on the draft Range ALDs for grades 5 and 7.) After consulting with adjacent-grade groups, participants returned to their grade level groups to continue their within-grade Range ALD development. A more structured cross-grade level articulation occurred in the other subgroup. Group members shared information on how they were writing their Range ALDs. In both cases, the purpose of these procedures was to promote cross-grade consistency of the final Range ALDs.

Participants spent the remainder of the afternoon completing the grade-level Range ALD creation task. An afternoon break occurred at roughly 2:30pm; one group went on break essentially as a group; the other group break was more informally structured, with group members breaking at times they deemed appropriate for their work. During the afternoon activities, participants were encouraged to keep cross-grade articulation in mind and various activities (some more formal full-group discussions/reports, some less formal cross-grade conversations) were incorporated to help facilitate homogeneity of the final draft Range ALDs across the grade levels covered by each subgroup. In whole-group conversations, participants commented on the benefits of having engaged in the Range ALD creation task and the cross-grade articulation.

One of the groups finished their work slightly ahead of schedule and were dismissed at approximately 4:00pm. The other group took slightly longer and completed their work by approximately 4:45pm. Groups were dismissed for the day after being reminded to keep any materials, written notes, etc., secure and to place them in their folders; they were reminded about some housekeeping details such as reimbursement requirements and the start time for Day 2; and participants were thanked for their Day 1 engagement. All materials were to be left in the room, collected, and returned to them in the morning.

Day 1 end with a DRC/NCDPI debriefing and planning meeting, similar to what had occurred over the lunch break. No major issues were identified; the results of each group's activities on Day 1 were summarized, and plans for Day 2 were reviewed.

Day 2: Morning Activities

The second day of standard setting activities began on Tuesday, July 9, 2019 with sign-in, materials pick-up, and refreshments for participants beginning at 7:30am. The day's work activities began at approximately 8:40am. The first part of the morning activities consisted of an orientation to the particular methods that would be used to derive cut scores to define the performance levels. Two different methods were used, one for the general assessment (the Bookmark method; Lewis, Mitzel, Mercado, & Schulz, 2012) and one for the alternate assessment program (Yes/No Angoff; Plake & Cizek, 2012). The Bookmark method requires participants to examine a booklet containing items in a test, printed one per page and ordered from easiest to most difficult. Participants then make judgements about the items that students just entering various performance levels ("threshold students") should be expected to answer correctly by placing an indicator ("bookmark") at the location in the booklet that separates the items they judge that threshold students have at least a 50% chance of answering correctly from those they believe that

the threshold students have less than a 50% chance of answering correctly. The Yes/No Angoff method also requires participants to make judgments about the performance of threshold students; using this method, participants judge whether threshold students have at least a 50% chance of earning 2 points (the maximum score) on an item, 1 point, or zero points. The steps for implementing both methods is similar, involving:

* review of NC test items,
* making judgments about what "threshold" students (i.e., students on the borderline between two performance categories) should know or be able to do,
* considering various sources of feedback and other information,
* engaging in group discussions and revising judgments, and
* after three rounds of judgments, arriving at final, group recommended cut scores.

After general descriptions of the methods to be used, hands-on practice in using the methods was provided. Two concerns arose in the group charged with setting standards on the alternate assessments. Those included concerns by participants about making judgments out of grade level expertise, student population expertise; and nervousness about making judgments about threshold student performance without more consideration of the characteristics of those threshold students. The concerns were addressed by the facilitator and via group discussions. By the end of the practice activity, participants appeared to be more comfortable with their judgment tasks.

At the conclusion of the practice activity, the group considering the alternate assessment took a short break (at approximately 10:00am); they then began an activity as a whole group, working in tables, developing the 2/3 and 3/4 threshold student descriptions for grade 6. The group considering the general assessment took a short break at approximately 9:45am; they then broke into three grade-level groups and, in separate rooms, began an activity developing the 2/3, 3/4, and

4/5 threshold student descriptions for grade 4, 7, and Math 1. The threshold description activity concluded with the tables at each grade level coming together for full grade-level group discussions of the characteristics each table group had identified as relevant to the threshold students. The purpose (and result) of this activity was to promote a common conception of the threshold students for all participants. These discussions concluded at approximately 11:30am.

The final activity on the morning of Day 2 was originally planned to be an opportunity for participants to encounter the test for their grade levels in the same way as students would experience that test; that is, participants would have an opportunity to take and self-score a version of the relevant grade-level test. The purpose of this activity was to provide participants with a concrete sense of the range of content covered by the test, the difficulty of items on the test, and the overall level of challenge presented to the students.

In the alternate assessment group, the group was provided with a copy of the grade 6 test form and directed to use the form to familiarize themselves with the test. This activity began at approximately 11:25am. The group was directed to review the test as much as they wished, with the understanding that they should complete the review activity by approximately noon when the lunch break would begin. It was also announced that participants identified as table leaders should remain in the room and not go immediately to lunch at the conclusion of their test review so that they could be given some specific instructions related to their table leader roles.

In the general assessment group, the original plan was for participants to self-administer a released test form. However, it was decided that many participants would likely have already seen the released form and that reviewing that form again would not provide an experience that would align with the intended purpose of the activity. Instead of a test-taking experience, participants were instructed to choose a few problems somewhat randomly in an ordered item booklet (OIB) and try to answer a few of the questions. (An OIB is commonly used in Bookmark standard setting

procedures; it typically consists of all items comprising a test form, presented one item per page, in order from easiest to most difficult item in the form.) After having this experience with a few items, participants were introduced to the task they would be doing next—independently reviewing all items in their OIBs and making annotations on an "item map" regarding the knowledge and skills that would be necessary for a student to answer that item correctly. The "item map" is a listing of items in a test form, ordered in the same way as the OIB, and indicating the standard measured by the item and other descriptive information about each item.

Overall, given that the intended original purpose of the test experience activity was to provide participants with a concrete sense of the range of the challenge presented to the students, it is unclear if the modified review procedure (compared to the actual taking of the test as originally planned) provided the same frame of reference for participants related to the level of challenge represented by the test. Nonetheless, it was clear that participants appreciated the opportunity to review the actual test that students had been administered; the modified review likely aided them in better understanding the items they would be reviewing and rating later in the standard setting workshop; and, in the Bookmark groups, the work reviewing the OIBs and completing the item maps also likely provided good background on the level of challenge represented by the test items.

Day 2: Afternoon Activities

Activities on the afternoon of the second day began at approximately 1:00pm. The activities differed for the Yes/No Angoff (i.e., alternate assessment) and Bookmark (i.e., general assessments) groups.

NCEXTEND1 (Alternate Assessment) Activities

The alternate assessment group was first asked to complete a process evaluation which sought their impressions of the training and readiness to proceed. After completing the evaluation,

the facilitator briefly reviewed the Yes/No Angoff procedures. Then, each participant was provided with a copy of a grade 6 test booklet and rating sheet and they were instructed to begin a first round of ratings for all items in the booklet, completing their Round 1 judgments independently by approximately 2:30pm. The group was instructed to complete their judgments first for all items with respect to the 2/3 threshold, then to cycle back through the test and provide judgments for all items with respect to the 3/4 threshold.

After an afternoon break that provided DRC staff with time to collect and summarize participants' judgments, Round 1 feedback was provided to the panelists. The feedback consisted of a graphic showing the distribution of all panelists' recommended raw cut scores for the 2/3 and 3/4 performance levels and the median recommended cut scores based on the total group; the same information was shown by tables. In addition, "benchmark" feedback data were provided, comprising the raw cut scores currently in place from the 2013 forms. Participants were encouraged to discuss this information in their table groups. After discussion, participants were directed to provide their Round 2 ratings for all items, considering the group discussions, the descriptions of the threshold students, and the content of the test items. The group finished making their Round 2 judgments by approximately 4:10pm at which time DRC staff began analysis of the Round 2 data.

At 4:40pm, Round 2 data analysis was completed and results were again presented to the group; the results were shown as whole group results and as table-level results, again using a graphic frequency distribution display to show the median and range of raw cut score judgments. Additionally, there was slightly less variation in the distribution of individual recommended cut scores. In addition, "impact" feedback data were provided at the end of Round 2 for the first time. Impact data show the percentage of students in the state who would be classified into each of the three performance level if the total group's Round 2 cut score recommendations were used.

An abbreviated discussion of these results occurred and participants were given the option of staying later in order to complete their Round 3 ratings, or to generate those judgments at the beginning of Day 3. The group decided to complete their Round 3 judgments, with an understanding that, due to the shortness of scheduled time remaining for the day, review and reconsideration of the Round 3 judgments may be needed. All group members finished that task, submitted their rating sheets, and turned in their secure materials by 4:45pm.

NC General Assessment Activities

The Day 2 afternoon activities for the general assessment group began with a continuation of their review of the OIBs in their grade-level breakout rooms. The groups were instructed to finish their task of analyzing each item in the OIB, making notes on their item maps regarding the characteristics of the items that contributed to their cognitive challenge for students, and to complete their OIB reviews by approximately 3:00pm. An afternoon break was taken at 3:00pm and, at 3:15pm, the grade-level groups came together as a whole group in a large conference room for the purpose of receiving specific, common training in the Bookmark procedures. Following training, the Bookmark group completed an evaluation similar to that which was completed by the Yes/No Angoff group. When the evaluation activity was finished, all participants left large-group conference room and reassembled in their grade-level (i.e., grades 4, 7, Math 1) break-out rooms. In the grade-level rooms, they were provided with grade-relevant OIBs, rating forms, and all materials necessary to generate their individual Round 1 Bookmark ratings. Participants made three threshold recommendations, beginning with the 3/4 bookmark placement, then considering the 4/5 bookmark placement, and ending with recommendations for the 2/3 bookmark placement.

At the conclusion of the first day, NCDPI Section Chief Maxey-Moore again convened a meeting of DRC and DPI staff; the purpose of the meeting was to allow facilitators to discuss any concerns, resolve any issues, and tie up any loose ends from the day's work so that the standard

setting activity could continue to progress smoothly into Day 3. No issues required attention related to the general assessment Bookmark activities.

A central issue discussed at the end of Day 2 was the substantial disparity between previous NCEXTEND1 impact and the impact that would result based on standard setting participants' current recommendations. Several potential adjustments that could be implemented in Day 3 were discussed. It was observed that, for many items, the panelists' judgments were not well aligned with the actual performance of North Carolina students on those items. It was decided to adjust the process to begin Day 3 with examples of actual NCEXTEND1 test items, the corresponding panelists' ratings for those items, and the actual performance of students in terms of the percentage of students who answered the items correctly (i.e., *p*-values). DRC staff worked to prepare materials to allow participants to consider this kind of information consisting of a list showing, for each item in the test, the actual performance of NC students. It was decided to begin Day 3 with an opportunity for participants to review their Round 3 judgments, having the benefit of that actual performance information.[1]

Day 3: Morning Activities

The third day of standard setting activities began on Wednesday, July 10, 2019 with sign-in, materials pick-up, and refreshments for participants beginning at 7:30am. The day's work activities began at approximately 8:35am. Prior to participants' activities, DRC and NCDPI met to confirm the adjustment in procedures for the NCEXTEND1 panel.

NCEXTEND1 Activities

The first activity for the alternate assessment group was a presentation by DRC on their

1 When this procedure was implemented on Day 3, it was observed that this information was considered to be helpful by the participants in making their judgments. Because of this, and because of a desire to implement consistent procedures for every grade level, a column of *p*-values was added to the rating sheet used by participants for generating all of their subsequent Round 3 judgments.

Round 3 results, information on how to use the additional information provided, and large group discussion about the other potential factors that could account for the differences in percentages in performance categories that were observed. These included whether to take student guessing into account when providing their judgments, differences in the specific match in wording between test items and the content standards, and the wordings of the ALDs that the groups had developed. Participants requested, and were provided with their Round 3 rating sheets in order to aid them in generating their next round of ratings. The final review of ratings began at approximately 9:10am;. Upon completing the final ratings, each participant drafted a written statement that provided a brief rationale for their cut score recommendations. The group completed all work for setting the grade 6 standards by approximately 10:00am. The group then took a morning break, before reconvening in separate grade-level groups to apply the same procedures they had used for grade 6 to determine their performance standards recommendations for grades 5 and 7.

The first activity for the two grade-level groups was to develop the Threshold ALDs for their respective grade levels. For one of the groups (grade 5), one of the initial activities included working through their grade level test; for the other group (grade 7), participants began directly working on their Threshold ALDs. Similar templates were used for the ALD development activities; grade-level groups worked at tables of 4-5 participants to develop their ALDs; each table group focused on a specific area (e.g., Number Sense, Ratio and Proportion, etc.). The groups then provided their Round 1 recommendations; all judgments were submitted prior to breaking for lunch at Noon.

NC General Assessment Activities

The first activity for the general assessment groups was provision of feedback on their Round 1 ratings. The work occurred in three grade-level groups. Each grade-level group was presented with histograms showing the frequency distributions of all individuals' ratings for the group and information on ratings for each table. Discussion of the results occurred at tables within the grade-level rooms; at the conclusion of the discussions, the groups were instructed to complete their Round 2 Bookmark judgments. After analysis, feedback was provided to participants regarding their Round 2 judgments, including impact data (i.e., the percentages of NC students who would be classified into the four academic performance levels if the panelists' median recommendations were used as the cut scores). In addition to the impact data, the participants were also provided with benchmark data (i.e., data that showed the historical percentages at each of the academic performance levels). Following discussion of these results (including consideration of the ALDs, the threshold descriptions, the content standards, and connections between the standards and the test items), participants were instructed to complete their Round 3 ratings. Nearly all participants completed their ratings before breaking for lunch at 12:00pm.

Day 3: Afternoon Activities

The first activity of the afternoon for the alternate assessment group was a review of the final results for their morning's grade 6 judgments. After brief discussion of the results, the group continued work in grade-level rooms to develop Threshold ALDs (that is, descriptions of threshold student knowledge and skill) and provide rounds of judgments for grades 5 and 7. Following each round, the groups received the same kind of feedback (frequency distributions, whole group and table medians, benchmark data and impact data) as had been done previously. Additionally, each group was shown information that illustrated the performance standard recommendations for the relevant grade spans. These presentations consisted of stacked bar charts showing the percentages

of students in each performance level across all grades. (Note: the data shown for grades 5, 6, and 7 were based on the participants' recommended cut scores at the conclusion of Round 2; the data on the graphs for the other grades was benchmark data from the previous year's test administration.)

For the remainder of the afternoon, the alternate assessment participants continued their work in grade level groups. Each group continued Threshold ALD development and discussions, rounds of Yes/No Angoff judgments, considerations of feedback, discussions, and review of ratings for grades 5 and 7. By the end of the day, the upper level group finished their Round 2 Yes/No judgments for the grade 8 assessment. The lower level group completed their Round 1 judgments for the grade 4 assessment.

The afternoon activities for the general assessment group mirrored those of the alternate assessment group (with the exception that the general assessment group continued to use the Bookmark standard setting method). The groups continued working in grade-level rooms to develop Threshold ALDs and provide rounds of Bookmark placements in grade-level rooms for grades 4, 7, and NC Math 1. Following each round, the groups received the same kind of feedback (frequency distributions, whole group and table medians, benchmark data and impact data) as had been done previously. Additionally, each group was shown information that illustrated how the performance standard recommendations for the relevant grades spans were aligned as had been done in the alternate assessment feedback sessions. For the remainder of the afternoon, the general assessment participants continued ALD development and rounds of Bookmark placements, considerations of feedback, discussions, and review of ratings for each grade level within their assigned grade bands.

By the end of Day 3, the general assessment groups had completed Round 1 Bookmark judgments for the grade 3 assessment; Round 2 Bookmark judgments for the Grade 6 assessment,

and the upper grades group began initial review of the OIB for the Math 3 assessment. provided.

At the end of the day, the same procedures were followed as had been done on previous days; these included log in of secure test documents, collection of participants' materials, and a reminder of the final day's agenda. Also, as had been done at the end of each previous day of the standard setting activities, a debriefing session was held, attended by contractor and NCDPI staff in order to share progress on Day 3 and make any needed adjustments in procedures for Day 4.

Day 4 Activities

The final day of standard setting activities began at approximately 7:30am on Thursday, July 11, 2019 with sign-in, materials pick-up, refreshments for participants, and assistance with processing of reimbursement. The day's work activities began at approximately 8:30am. The general and alternate assessment groups continued work on their assigned tests; the alternate assessment groups finished all rounds of ratings for their final assessment by approximately 2:00pm, with the general assessment groups completing all of their judgments shortly thereafter. When the groups were finished, they were thanked for their participation by Ms. Maxey-Moore, who also described the next steps involving their recommendations and who reminded participants about what information they were encouraged to share with colleagues in the field and what information must remain secure and confidential. Both groups completed an end-of-process evaluation, submitted all secure materials, reviewed honoraria and expense reimbursement procedures, and received continuing education credit certificates. Most participants were then dismissed; table leaders from all groups remained for the later afternoon vertical articulation activities.

Vertical articulation involves standard setting participant representatives from individual

grade levels (i.e., designated table leaders) assembling at the conclusion of setting standards at individual grades for the purpose of considering *all* of the recommended standards across grades as a coherent system of standards. Participants in vertical articulation are charged with smoothing out inconsistencies across all grades, with particular attention to historical trends and potential anomalies in current impacts.

Vertical Articulation – Alternate Assessments

At approximately 2:30pm on Day 4, vertical articulation sessions began for the purpose of finalizing recommendations for performance standards for the alternate assessments. The first step in the process was to assemble participants from the grade-band groups and present them with the impact data for grades 3-8 and NC Math 1 for the NCEXTEDN1 cut score recommendations. The purpose of this meeting was to assist table leaders in being able to represent their groups' perspectives when the table leaders meet for the second step of vertical articulation later in the afternoon. Impact data were presented as a graphic (stacked bar charts) that showed each grade and each performance level. Similar data showing historic impact rates for all grades and performance levels was also shown. The facilitator explained the task of promoting consistency across grade levels, although it was not intended to promote any specific result or sameness of results across the grades. The group considered the data and a thoughtful discussion ensued regarding which individual grade results seemed most appropriate and which seemed most warranting of adjustment.

For the second step of the alternate assessment articulation process, table leaders representing all grade levels were assembled, presented with the same graphical information, and given the same directions as to the purpose of the articulation activity. After discussion of the results and benchmarks, group members were asked about potential changes to the overall system of the results that would introduce greater consistency and coherence. The actual system for

processing this information seemed somewhat crude and there was some confusion about actually generating the results. For a single score point change at a single performance level at one grade, it appeared as if recommended changes had to be manually entered, an excel spreadsheet updated, and a new graph produced; the process seemed to take an inordinate amount of time and broke the flow of panelists' discussions.

Eventually, the articulation projection worked more smoothly, and the group made minor adjustments to three grade levels to accomplish a more coherent system of impact that seemed reasonable and aligned to the content demands of each grade's ALDs. The group arrived at a final consensus recommendation and were dismissed at approximately 4:00pm.

Vertical Articulation – General Assessments

At approximately 4:10pm, vertical articulation for the general assessment began. The process was conducted slightly differently than for the alternate assessment in three ways: 1) general assessment grade-level groups were not individually presented with results across all grade levels; 2) the articulation session involving table leaders from each grade group were first presented with a graphic showing historical impact over several recent years; and 3) potential changes in the system of cut scores were initially introduced by the facilitator (within the range of 1-2 raw score point changes at three of the grade levels) as opposed to initial suggestions for grade levels and magnitudes of changes being suggested by group members.

The vertical articulation session continued with presentation of stacked bar charts showing the percentages of students that would be classified into each performance level based on whole group recommended cut scores (medians) that were obtained following the final (i.e., Round 3) judgments. The facilitator then initiated conversation among the table leaders regarding their reactions to the initial results. Group comments focused on the extent of students who would need support to be successful in the next grade, the rigor of the content students were exposed to at each

grade, the unique nature of the grade 8 and NC Math 1 assessment populations, and the challenge presented by the items in the OIBs.

An initial question for the group focused on the reasonableness of the impact of the Level 5 panelist recommended cut scores. The group appeared to broadly endorse the panelists' recommended impact. Next, modest (e.g., 1-2 raw score point) changes for selected grade levels that would produce a more consistent system of cut scores across the grades, and table leader reactions to those changes were solicited and discussed. In the interest of time, the facilitator summarized some of the themes that had been expressed in the discussions and in participants' conversations over the course of the week, and projected a system of adjusted cut scores that attempted to maintain fidelity to the range of participants' intentions, the content expectations of the NC general assessments, and historical impact patterns.

III. Summary and Recommendations

Based on my observations of the procedures, materials, and processes used to obtain recommended performance standards, it is my opinion that the standard setting activities implemented for the North Carolina general and alternate (NCEXTEND1) mathematics assessments were, overall, conducted in a manner consistent with sound psychometric practices. The resulting panelist recommendations can be viewed as valid estimates of appropriate cut scores for the North Carolina assessment program.

Overall, the process was characterized by a number of strengths; few concerns arose during the course of the standard setting. In the following sections some key strengths and suggestions for the future are described.

Strengths

1) The contractor for setting performance standards on the North Carolina mathematics examinations developed appropriate and reasonably specific plans for implementing accepted standard setting methods (i.e., the Bookmark and Yes/No Angoff methods).

2) The state's technical advisors reviewed key elements of the plans in advance and judged them to be sound and defensible.

3) Overall, the implementation of the plans appeared to be well organized and a faithful implementation of the standard Bookmark and Yes/No Angoff procedures. The contractor provided adequate resources and personnel to ensure that the standard setting was conducted professionally and paced appropriately. The relevant experience included expertise in psychometrics and mathematics content expertise. The content specialists who supported the whole-group and small-group breakout sessions were knowledgeable about the North Carolina content standards and ALDs, non-intrusive, and they provided clear

guidance to the participants.

4) Participants in the standard setting activities had relevant qualifications for making the judgments they were asked to make. Participants consisted of North Carolina educators with experience teaching mathematics at relevant grade levels, and educators having experience with either general population students, special needs students, or both. All participants appeared to be motivated to complete their work conscientiously, and they worked attentively and thoughtfully. No issues regarding personal agendas or domination of discussion in groups/tables were apparent.

5) Participants appeared to understand the standard setting tasks they were to perform, and the nature of the feedback provided to them (i.e., normative and impact information). Participants who were identified as table leaders functioned well in their roles.

6) Technology used to support the standard setting activities (e.g., laptop computers, PowerPoint projections, audio equipment, scanners, etc.) functioned well.

7) The meeting arrangements, food service, and other logistics appeared to meet the needs of the participants. Contractor staff were highly attentive to ensuring that meeting rooms were comfortable and conducive to supporting participants' work.

8) The materials, forms, and other items used appeared to be well-designed and easy for participants to use.

9) There was appropriate concern for and attention to confidentiality and security of materials and results.

Conclusions

Four summaries seem warranted from the data available and the observations conducted of the current standard setting workshop:

1) Performance standards for the NC general and alternate mathematics assessments were

recommended by well-qualified, engaged, and thoughtful groups of North Carolina educators. The standard setting plan was developed and implemented by qualified and conscientious contractor staff. The entire endeavor was overseen by qualified, attentive, and experienced NCDPI staff.

2) Overall, few issues of concern arose during the standard setting process for the North Carolina assessments. Most issues were relatively minor and did not appear to have immediate, discernible effects on the procedures or results. When issues of substance arose, they were identified and discussed by contractor and NCDPI personnel. Adjustments to the intended procedures were reasonable, appropriate, and supported the integrity of the standard setting process. Minor suggestions for improvement that NCDPI might consider for future standard setting endeavors are noted previously in this report.

3) The procedures implemented for recommending performance standards on the North Carolina mathematics assessments generally followed best practices for standard setting and were generally faithful to the specific methodological procedures intended (i.e., Yes/No Angoff, Bookmark). Note: one important source of information was not available at the time this report was written: the results of participants' evaluations. NCDPI should review the evaluation results in order to confirm or qualify the conclusions of this report.

4) The vertical articulation procedures provided an effective mechanism for participants (through their table leaders) to address fluctuations in results in a way that produced a more coherent and consistent system of performance standards.

The preceding four summaries support the conclusion that the participants' cut score recommendations can be considered to be valid and reliable estimates of the cut scores for the relevant assessments. Unless the panelists' evaluations indicate otherwise, policy makers can have confidence that the recommendations from the standard setting activity are based on sound procedures, producing credible, defensible, and educationally useful results.

IV. References

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

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

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.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (pp. 181-200) New York: Routledge.