Observation and Report on

North Carolina English II

Standard Setting Workshops

August 4-6, 2020

Prepared by:

Gregory J. Cizek, PhD

August 19, 2020

Executive Summary

The North Carolina Department of Public Instruction (NCDPI) contracted with Data Recognition Corporation (DRC) to engage in a process called *standard setting*, which comprises procedures for deriving performance standards (aka, "academic achievement levels," or "cut scores") for tests. In this case, a standard setting process was developed and implemented for setting performance standards for the North Carolina English II end-of-course (EOC) student achievement test. Three performance standards were identified to define four levels (*Level 5, Level 4, Level 3* and *Not Proficient*).

A large, diverse, and representative group of qualified participants—namely, North Carolina classroom educators—were empaneled to perform the standard setting procedures on August 4-6, 2020. All standard setting activities were conducted through the use of a virtual meeting platform. A total of seven educators participated in the first major activity of the English II standard setting activities—achievement level descriptor development—and 14 educators (five of whom had participated in the previous day's achievement level development session) served as panelists for the second major activity—the actual standard setting process. It should be noted that, although the phrase "standard setting" is used in this document, it is most accurate to note that the panel did not actually *set* the performance standards. Rather, because the North Carolina State Board of Education (NCSBE) is the entity with the authority to establish performance standards on NC state assessments, the process is most accurately described as "standard-*recommending.*" Results from the standard setting activities described here are subsequently reviewed by relevant personnel at NCDPI and potentially approved by the NCSBE.

Overall, the standard setting activities comprised two related tasks: 1) a review and refinement of the Achievement Level Descriptions (ALDs) that define performance at each of the

performance levels used to report student achievement on the English II assessment (i.e., Level 5, Level 4, Level 3, Not Proficient), and review and judgments regarding the level of achievement required for students to be classified into each of those levels. A process relying on expert review was used for ALD development; the Bookmark procedure (Lewis, Mitzel, Mercado, & Schulz, 2012) was used obtaining recommended performance standards.

Qualified educators from North Carolina were trained in these procedures; they were facilitated in the session by content and process specialists from a testing contractor, DRC and observed by staff from NCDPI. Judgments regarding cut scores for the English II assessment were obtained based on subject area experts' content-based judgments across three rounds of judgments.

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) and has served as an independent, external observer of previous standard setting activities for other NC assessments. The author also serves in an on-going basis as an NCDPI technical advisor.

Overall, the workshop produced well-articulated ALDs and cut score recommendations that can be considered to be valid and reliable estimates of appropriate performance standards for the English II assessment. 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) Workshop Observations; 3) Conclusions and Summary; and 4) References.

II. Workshop Observations

Day 1: Morning Activities

All activities for the English II EOC standard setting were conducted virtually using a teleconferencing platform. The morning sessions on Tuesday, August 4, 2020 began at approximately 8:30 a.m., with a welcome to participants provided by the facilitator (Dr. Rick Mercado, DRC) and NCDPI leadership (Ms. Kristen Maxey-Moore, Section Chief, Test Development). Dr. Mercado provided information on some of the functionalities of the virtual meeting site that participants might find helpful; he also reviewed the pre-meeting materials that were provided to participants; and he directed participants to complete a pre-meeting survey. Ms. Maxey-Moore introduced other NCDPI staff who would be participating in the workshop. She then projected a PowerPoint presentation that provided participants with some fundamental testing concepts, an overview of the NCDPI test development process, information on how participants were selected for the standard setting workshop, and a brief description of the timeline for the new English II assessment and where standard setting activities fit into that timeline. Ms. Maxey-Moore ended her presentation inviting participants to ask any questions about the process and thanking them again for their participation.

At approximately 9:15 a.m., participants were dismissed for a short break before being assigned to virtual breakout rooms. A total of seven educators participated in the first day of English II standard setting activities which focused on review and development of achievement level descriptions (ALDs) that describe the kind of student performance intended to be captured at each of the performance levels (i.e., Level 5, Level 4, Level 3, Not Proficient). At approximately 9:20 a.m., Dr. Anne Kirpes (DRC content facilitator), oriented participants to this activity by

providing an overview of the different kinds of achievement level descriptions: Policy ALDs, Range ALDs, Threshold ALDs, and Reporting ALDs. Dr. Kirpes also introduced participants to a central electronic depository of workshop materials (called the "Workshop HUB"), which contained copies of the workshop agenda, the NC content standards, draft ALDs, and workshop surveys and evaluations.

Participants used a draft of the Range ALDs developed by content specialists at DRC and NCDPI. The first morning task for participants was to independently review the draft ALDs. Beginning with the Grade 8 Reading content standards, participants were asked to review the progression of knowledge and skill and changes in cognitive complexity, as the content standards express the range of student performance from Not Proficient to Level 5. Participants were directed to begin their review with attention to the Range ALDs for Level 4, and to note the appropriateness of the Level 4 ALDs, including the extent to which those ALDs captured expectations that were in line with the Policy ALDs, whether they were missing any key elements, went beyond expectations, etc. Participants were also asked to then review the draft descriptions of the other achievement levels and comment on whether they captured appropriate progressions of knowledge and skill, increases in cognitive complexity across levels, relevant key elements, and so on. Participants completed this activity independently by approximately 10:15 a.m.

Beginning with a single, specific content standard, Dr. Kirpes then facilitated a review of group-level consideration of participants' observations related to the content standards and achievement level descriptions. Each participant took the lead on discussing an individual standard. At 11:45 a.m., discussion of grade 8 ALDs wrapped up and participants were directed to begin their independent reviews of the ALDs for the upper grade levels (i.e., grades 9-10). Participants were informed that a lunch break was scheduled for noon; however, if needed,

additional time could be provided after the break to complete the independent reviews. Panelists then took a lunch break from 12:00-12:45 p.m.

Overall, the Day 1 morning activities proceeded smoothly. Perhaps due to limitations of internet connectivity, one participant had some challenges with the video portion of the ALD review but was able to fully participate in the discussions. The virtual meeting platform appeared to work well for the ALD review; the session was well-facilitated; and all panelists appeared to engage thoughtfully and productively in the ALD review process.

Day 1: Afternoon Activities

After returning from lunch at 12:45 p.m., panelists indicated that they needed a few additional minutes to complete their independent ALD reviews. All panelists completed this activity and the Dr. Kirpes facilitated a group discussion that began at 1:00 p.m. and concluded at approximately 1:55 p. m. An afternoon break was taken from 1:55-2:10 p.m. At the conclusion of the break, the facilitator led an articulation review across grade levels; that is, panelists were asked to look at the Level 4 ALDs for a content standard across grades 7, 8, and 9-10 to ascertain the degree to which an increasingly level of demand/complexity was present going up the grade levels. The group engaged in this activity independently from approximately 2:15-2:30 p.m., followed by discussion of this review until approximately 2:45 p.m. At the conclusion of the discussion, panelists engaged in independent reviews of the ALDs for other performance levels, focusing on the Level 3 ALDs; a group discussion of this review was facilitated by Dr. Kirpes beginning at approximately 3:00 p.m. The afternoon ALD review session ended at approximately 4:10 p.m. Panelists were then thanked for their participation by Dr. Kirpes and Ms. Maxey-Moore. Ms. Maxey-Moore emphasized the role that panelists can play in being ambassadors for the process;

how ALDs will be used in standard setting; and what the next steps in the process are (i.e., standard setting activities on Wednesday and Thursday, recommended standards to the NCSBE in September; and English II scores and performance levels reported in October). She asked if panelists had any questions (there were none) and panelists were reminded to complete the workshop evaluation in the Hub. Finally, Sara Kendallen (DRC) reminded panelists about security of materials and procedures for returning secure materials to DRC.

After all panelists had left the meeting, a debriefing session among DRC, NCDPI, and the observer began at approximately 4:20 p.m. A review of the day's activities indicated no concerns that would require attention. The focus of the debriefing session was identification of potential breakout room leaders for the standard setting activities that would take place on Wednesday and Thursday, and configuration of the threshold description activity. Dr. Mercado reviewed the specific activities planned for the next day so that everyone was on the same page going forward. The debriefing activity concluded at approximately 5:00 p.m.

Day 2: Morning Activities

To begin the second day of standard setting activities on Wednesday, August 5, 2020, participants entered the virtual meeting platform starting at approximately 8:00 a.m. and were welcomed by DRC program manager, Julie Korts. Ms. Korts alerted each participant to a pre-meeting survey that was accessible via the Zoom chat. All participants completed the survey and were ready to begin the meeting at the scheduled time. A total of 14 panelists participated in the English II standard setting workshop, including five panelists who had participated in the ALD meeting on the previous day.

At approximately 8:40 a.m., Dr. Mercado welcomed participants and alerted selected

panelists to a chat message that invited them to serve as breakout room leaders for the standard setting activities. Dr. Mercado also provided a few tips on using the virtual meeting platform and Ms. Korts gave participants information on materials that were included in the packets mailed to them prior to the meeting, how to return materials to DRC, and reminders regarding what information from the meeting could be shared and what information should remain secure. Shortly thereafter, Ms. Maxey-Moore officially opened the session, again welcoming the participants and thanking them for their commitment to the important tasks ahead. Ms. Maxey-Moore first led introductions of NCDPI staff, then provided an in-depth presentation on the development process for North Carolina assessments; the presentation was the same as provided to participants in the ALD development session held on the previous day and included description of each of the steps followed for NC test development, the timelines for test development, and the need for and function of standard setting.

At approximately 9:10 a.m., Dr. Mercado facilitated introductions of DRC staff and panelists. The panelists all indicated relevant expertise and experience for English II standard setting and appeared to be diverse in terms of representing all areas of the state. At the conclusion of the introductions, Dr. Mercado provided an overview of all workshop activities, beginning with answering the question "What is standard setting?" He then gave presentation that included information on ALDs, "threshold" students on the borderline of achievement levels, the bookmark procedure (Lewis, Mitzel, Mercado, & Schulz, 2012) that would be used for standard setting, the ordered item booklets (OIB) and item maps used in that procedure, and instructions on how to record their bookmark judgments. At approximately 10:00 a.m., Dr. Mercado led the group in some practice activities, including a discussion of the threshold student that addressed the characteristics of such a student vis a vis a specific content standard, consideration of the

knowledge and skills required for answering specific items in an OIB (including both dichotomous and polytomously-scored items), and examples of feedback that participants will be receiving between rounds of judgments. A break in the morning activities occurred between 10:30 and 10:45 a.m.

At 10:45 a.m. Dr. Mercado introduced the "Hub" electronic repository of materials that would be used in the standard setting workshop. He then described the primary morning task for which participants would separate into three breakout rooms, with each breakout room focusing on developing a description of one of the threshold students (i.e., one room focused on the Not Proficient/Level 3 threshold student; one focused on the Level 3/Level 4 threshold student; and one focused on the Level 4/Level 5 threshold student). The assigned task required participants to review the content standards and ALDs, then to describe the knowledge and skills possessed by students at their assigned threshold. Participants completed that task and were dismissed for a lunch break at Noon.

Day 2: Afternoon Activities

Panelists were dismissed for a lunch break at Noon, and returned at approximately 12:30 p.m. The first activity after lunch involved a whole group discussion in which each of the threshold break-out groups reported on their conceptualizations of their threshold level students facilitated by Dr. Kirpes. The whole group discussion ended at approximately 1:15 p.m. The next activity was an opportunity for panelists to become familiar with the English II test content by independently reviewing a practice test. The purpose of this exposure to a sample of actual test items is to aid panelists in better understanding the content standards, seeing how they are operationalized in an assessment, understanding the level of challenge required of students, the

different item formats, and the different levels of cognitive complexity represented by the items. Using materials securely provided via the DRC Hub, panelists completed this activity by 1:45 p.m.

Beginning at approximately 1:50 p.m., an introduction for the next activity was provided by Dr. Mercado who demonstrated how to use the Hub materials, the item map, the 73-page OIB and the standard setting platform. Participants used these materials analyze the cognitive demands of items in the OIB and to make annotations—in particular, annotations regarding the demands of the items with respect to the threshold students. Because this is considered a fundamental aspect of the Bookmark procedure and because of the rigor of the task, the remainder of the afternoon agenda was devoted to this activity. Breakout room facilitators officially "checked in" with participants roughly every 30 minutes, or as needed in response to participants' questions. It was anticipated that panelists would complete a review of approximately 50 items over the rest of the afternoon session, with time on the next day's morning agenda to complete the review of the entire OIB.

At the end of the day, the breakout room facilitators addressed their groups to get their reactions to going through the OIB and the item mapping, and to give them instructions for maintaining the security of materials overnight. Following this session, panelists were dismissed for the day at 4:30 p.m.

As had occurred at the end of Day 1, after all panelists had left the meeting, a debriefing session was held among DRC staff, NCDPI staff, and the observer. A review of the day's activities indicated no issues of concern and all personnel reported that the process was progressing smoothly. One minor agenda modification was announced: in response to some participants requests for additional time to do the item mapping/OIB review activity, it was decided to allow participants access to the secure materials at 8:00 a.m. the next morning (ahead of the official 8:30

a.m. start time) and to increase the allocated time for completing the activity in the next day's agenda. This minor deviation seemed appropriate—indeed, necessary and helpful in terms of the validity of the process—and unlikely to have any adverse consequences. The debriefing activity concluded at approximately 5:10 p.m.

Day 3: Morning Activities

The final day of activities began at approximately 8:00 a.m. (earlier than the originally planned time) to allow additional time for the OIB/item mapping review activity for participants who wished to do so. It was announced to panelists that there would be a hard stop for their review of 10:00 a.m. in order to leave adequate time for training in the actual procedures that would be used for determining the recommended cut scores.

At 10:00 a.m., all participants returned from their breakout rooms and gathered as a whole panel to receive instruction, led by Dr. Mercado, in using the item maps, their descriptions of threshold students and the OIBs to make their first round of independent bookmarking judgments. Panelists were reminded to ground their judgments in the threshold ALDs they had developed, and to begin their work by placing the Level 3/Level 4 bookmark, proceeding then to the Level 4/Level 5 bookmark, and finally to consider the Not Proficient/Level 3 bookmark. At the conclusion of Dr. Mercado's presentation, Dr. Kendallen directed participants to the DRC Hub to access and complete a mid-process evaluation. The purpose of the mid-process evaluation was to aid participants in self-assessing their understanding of the bookmarking process and to gauge their readiness to begin making actual cut score recommendations using the bookmarking procedure. Upon completion of the evaluation at approximately 11:05 a.m., Dr. Kendallen reviewed the survey items with participants as a group. After the review, and following a brief morning break,

participants were returned to their breakout rooms to begin making their first round of judgments for the English II assessment at approximately 11:15 a.m. Panelists were dismissed for a lunch break at Noon.

Day 3: Afternoon Activities

The final session of the standard setting began with two concurrent activities. After participants returned from the lunch break, Dr. Mercado led a whole-group review of the morning activities and prepared participants to complete their Round 1 bookmark judgments. At the same time, DRC personnel were analyzing the Round 1 judgments and, from approximately 12:45-1:10 p. m., the Round 1 results were presented to NCDPI personnel and discussed. The results included distributions of panelists' bookmark page placements, range and median placements, benchmark locations (i.e., page numbers corresponding to cut scores used on the previous form and corresponding to relevant ACT college and career benchmarks in reading and English), and impact (i.e., percentages of students that would be classified into each performance level if the Round 1 median recommendations were used). Results were provided for the group as a whole and separately by breakout groups.

With the exception of the impact data, the same results were then presented to the whole group of panelists. A whole-group discussion and reflection on the results occurred until approximately 1:30 p.m., when participants returned to their breakout rooms to continue the discussion in greater depth and to generate a second round of judgments. Breakout rooms were instructed to complete these tasks by 2:15 p.m., however some additional time was needed by panelists to complete the task.

When all data were submitted by panelists, they were analyzed by DRC staff and presented

to NCDPI staff at approximately 2:50 p.m. As was presented following Round 1, the results included distributions of panelists' bookmark page placements, range and median placements, benchmark locations (i.e., page numbers corresponding to cut scores used on the previous form and corresponding to relevant ACT college and career benchmarks in reading and English), and impact data (i.e., percentages of students that would be classified into each performance level if the Round 1 median recommendations were used).

All results seemed reasonable and in line with expectations for participants considering the information they were presented and their content-based judgments. A whole-group session to discuss panelists' Round 2 experiences began at approximately 3:00 p.m. During this whole-group activity, all of the information presented to NCDPI staff were also shared with participants, including impact data. At the conclusion of the discussions, participants were directed to generate their third, and final, round of ratings. Round 3 ratings were generated in the whole-group setting (i.e., not in the break-out rooms) and began at approximately 3:55 p.m. Participants completed their final rating task by 4:15 p.m. At that time, Dr. Mercado and Ms. Korts of DRC reminded participants of the importance of maintaining security of materials, procedures of returning materials to DRC, the process for requesting daily stipends, guidelines for sharing with others outside the standard setting sessions about the process used (but not the results), and what the next steps would be regarding their overall cut score recommendations.

When data analyses were completed, all results were again provided to the whole group of panelists (i.e., distributions of panelists' Round 3 bookmark page placements, range and median placements, benchmark locations, and impact data. After brief discussion of the final results, Dr. Mercado thanked the participants for their conscientious work over the past days and he directed participants to complete a post-workshop survey that was posted for participants in the DRC Hub.

To close the session, Ms. Moore also thanked participants on behalf of NCDPI. Participants then finished completing their post-workshop surveys and the meeting was adjourned.

III. Conclusions and Summary

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 English II EOC assessment were 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 nine key conclusions about the process are summarized.

1) The contractor for setting performance standards on the North Carolina English II developed appropriate plans for implementing the ALD development process and for obtaining performance standard recommendations.

2) Implementation of the plans appeared to be well organized and a faithful implementation of professionally-defensible procedures. It is particularly noteworthy that the contractor had initially planned to conduct the standard setting activities as an in-person workshop. However, because of the pandemic, it was necessary to re-tool the entire process in a very short timeline in order to conduct the activities virtually. Given that it had not been anticipated to conduct the workshop on-line, a substantial amount of restructuring and technology development was needed with little time to design, test, and refine the essential components.

3) Despite these obstacles, all aspects of the reconfigured system functioned essentially

flawlessly. In addition, it was perhaps especially fortuitous that all of the individual participants' personal technology appeared to function without major concerns as well, given that the workshop activities were conducted during the week of a hurricane that struck eastern North Carolina. (The hurricane did have the negative consequence of precluding NCDPI from conducting planned standard setting activities related to the NCEXTEND1 assessment.)

4) The contractor provided adequate resources and experienced and qualified personnel to ensure that the standard setting was conducted professionally and paced appropriately. The relevant experience included expertise in psychometrics and English language arts 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.

5) 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 a wide range of reading-related content at relevant grade levels (from 8th grade English to high level AP courses); the group comprised ethnic, gender, and regional diversity; and panelists had experience teaching a broad range of students. 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.

6) 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).

7) Participants who were identified as break-out room leaders functioned well in their roles.

8) The materials, forms, and platform developed by the contractor all 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.

Summary

Three summaries are warranted from the data available and the observations conducted of the English II EOC standard setting workshop:

1) ALDs and performance standards for the English II EOC assessment were recommended by a well-qualified, engaged, and thoughtful groups of North Carolina educators. The plan for deriving the ALDs and performance standards was developed and implemented by qualified and conscientious contractor staff. The entire endeavor was overseen by qualified and experienced NCDPI staff.

2) Overall, no issues of concern arose during the standard setting process. No events occurred that would weaken confidence in the validity of the panelists' recommendations.

3) The procedures throughout the sessions followed best practices for standard setting and were faithful to both the published methodological procedures and the plan designed by DRC and approved by NCDPI. (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.)

The preceding points support the overall conclusion that primary product of the standard setting effort—that is, 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.