DYNAMIC®
LEARNING MAPS

2016 Standard Setting: Science

Technical Report #16-03

10/24/2016

All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission provided the source is cited as:

Nash, B., Clark, A., Karvonen, M., & Brussow, J. (2016). *2016 standard setting: Science* (Technical Report No. 16-03). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Contents

List of Figures

List of Tables

Executive Summary

The Dynamic Learning Map™ Science Alternate Assessment standard setting event was conducted from June 15 – 17, 2016, in Kansas City, Missouri, following the first operational testing year in science. The standard setting was a Dynamic Learning Maps (DLM®) science consortium-wide event with the purpose of specifying a set of recommended cut points for the consortium's science assessment.

Panels consisting of representatives from partner states convened to recommend cut points. Separate panels were formed for fourth grade and fifth grade, which are assessed with the 3-5 grade band assessment; sixth grade and eighth grade, which are assessed with the 6-8 grade band assessment; the high school grade band; and the Biology course. Because expectations for students in lower grades within a grade span could reasonably be lower than expectations for students at higher grades within the same span, grade-specific achievement standards were needed for the lower grades. Three cut points were determined by each panel to differentiate between four performance levels.

A standard setting approach was implemented to classify student performance into four different levels: emerging, approaching the target, at target, and advanced. The approach was derived from existing methods, including generalized holistic and body of work, but modified to fit DLM's assessment design and scoring system. For DLM, the standard setting approach leveraged the linkage levels (i.e., levels of complexity) within each Essential Element (i.e., content standards) and the statistical modeling approach for determining student mastery classifications. DLM used a student profile approach to classify student mastery into performance levels. Profiles provided a holistic view of student performance across the Essential Elements and linkage levels. Cut points were determined by evaluating the total number of linkage levels mastered, similar to assigning a cut point along a scale score continuum.

Student profiles were developed to show student mastery (mastered/not mastered) for each of the three linkage levels for each Essential Element. There were two steps to determine overall student mastery. The first step used criteria for determining linkage level mastery classifications based on students' item responses. The second step was to calculate total numbers of linkage levels mastered in the subject. Profiles were then selected based on these values to be used as exemplars for standard setting.

Panelists were recruited to participate in the standard setting event from DLM partner states participating in the science assessment across all assessed grade levels. The majority of panelists were educators with experience in science and/or in teaching students with significant cognitive disabilities. Once panel selections were complete, panelists completed an online training module before the on-site standard setting event. This training provided a general overview of the DLM assessment system and was supplemented by additional on-site training on the standard setting panel procedures. Once on site, panelists were familiarized with the standard setting materials and methods, and then were given folders containing exemplars of student profiles to practice the rating process.

The standard setting process followed two basic steps: range finding and pinpointing. The purpose of range finding was for panelists to assign general divisions between performance levels after reviewing a limited set of exemplar profiles. After panelists determined the ranges of profiles where cut points were likely to be found, they moved on to the pinpointing process. During pinpointing, additional profiles were provided at levels within the range determined from the range-finding process. The purpose of pinpointing was for panelists to evaluate the additional exemplar profiles and hone in on specific cut points to distinguish the four performance levels. Within the range-finding and pinpointing phases, panelists had multiple opportunities to make independent evaluations. Further, at the end of the meeting panelists were asked to provide feedback as to their confidence with their group's recommended cut points and independently indicate a final recommended cut point if they were dissatisfied with the group's results.

By the end of the standard setting event, all panel-recommended cut points had successfully been identified. In all instances, the median individual recommended cut points and the group recommended cut point were the same. This suggests that overall the group process was effective for using expert judgment to classify student profiles into the DLM performance levels and identify corresponding cut points. Furthermore, a member of the DLM Technical Advisory Committee (TAC) was on-site for the standard setting event and reported back to the TAC on the overall quality of the event. Evaluations of panelists' experience with DLM standard setting as well as DLM TAC members' review of processes, outcomes and feedback from the observing member provide further evidence that the methods and process used were effective for achieving the goals of the meeting.

Following the panelist process, a statistical adjustment technique was applied to reduce the impact of panelist sampling on the cut points. Impact data was used to evaluate the distributions of students in each performance level category, with and without the statistical adjustments. The adjusted cut points and impact data across all grade levels were then presented to a vertical articulation panel convened during the standard setting event. The panel used content-based rationales to recommend that the statistically adjusted cuts be accepted for all cut points except for the grade 6 Emerging/Approaching cut point. This was the only cut point that increased as a result of the adjustment and the panel recommended retaining the non-adjusted lower cut point. The vertical articulation panel recommendation was accepted as the DLM staff recommended cut points. The DLM TAC and science state partners reviewed the panel recommended cut points as well as the DLM staff recommended cut points. After review, the TAC provided support for the statistical adjustment technique and overall standard setting process, and the state partners accepted the DLM staff recommended cut points.

The final set of cut points and impact data follow.

Table 1. DLM Recommended Cut Points for Science

Assessment Band	Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Maximum Number of Linkage Levels
3-5	4	9	15	21	27
3-5	5	10	17	25	27
6-8	6	9	15	21	27
6-8	8	10	16	23	27
HS	9-12	8	16	23	27
HS Bio	Biology	9	15	22	30

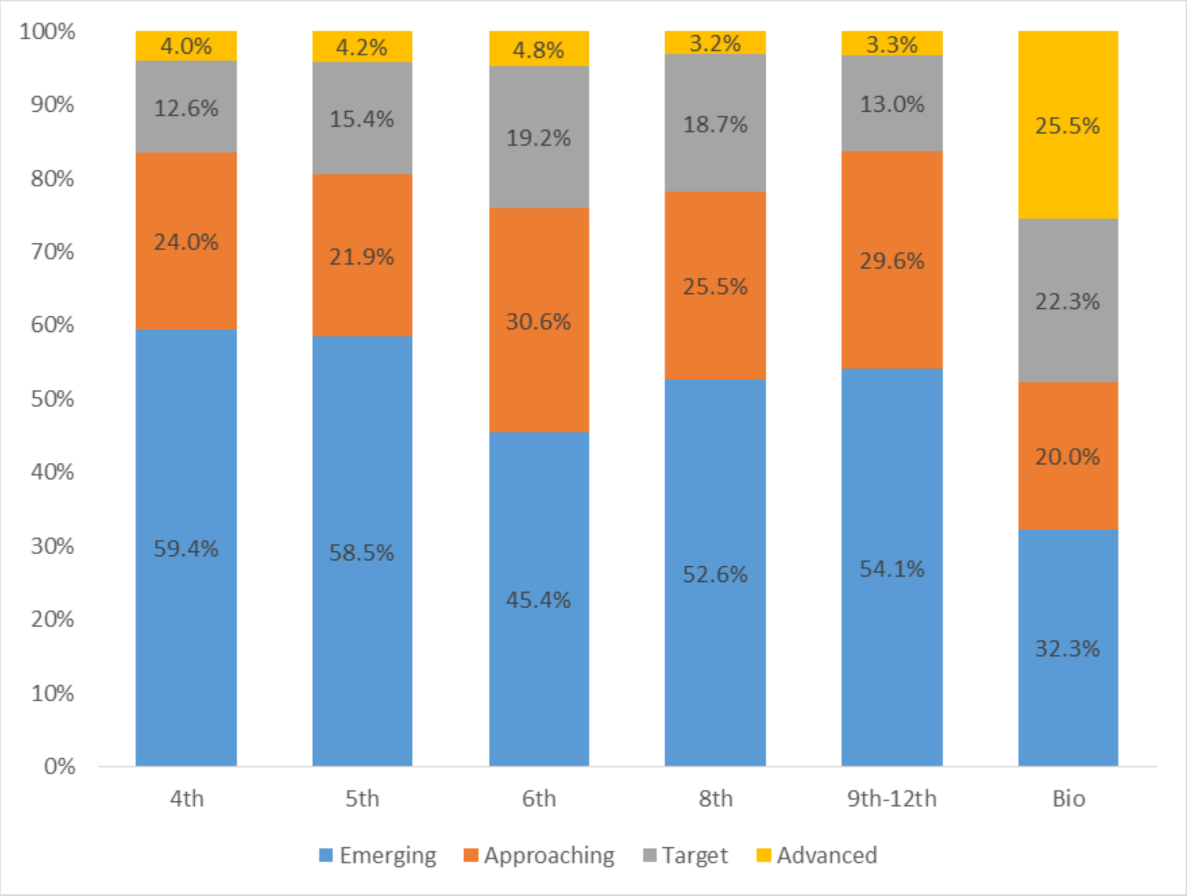


Figure 1. Impact Data Using DLM Recommended Cut Points for Science

Chapter 1: Introduction

The standard-setting process for the DLM science assessment consisted of the adoption of the existing DLM performance-level descriptors by the science states, a three-day standard-setting meeting, and follow-up evaluation of impact data and cut points by the state partners. This report provides an overview of the DLM assessment system and details the methods, preparation, procedures, and results of the science standard-setting meeting, including the follow-up evaluation of the impact data and cut points.

The purpose of the standard-setting activities was to derive recommended cut points for placing students into four performance levels based on results from the 2015-16 DLM science assessment. The intended audience for this standard-setting technical report is the DLM TAC, DLM state partners' state boards of education, and federal peer review committee members.

The 2015-2016 school year was the first operational testing year for DLM science assessments. The consortium operational testing window ended on June 10, 2016, and standard setting was conducted from June 15 – 17, 2016, in Kansas City, Missouri. The standard-setting event was a DLM consortium-wide event with the purpose of establishing a set of cut points for the science assessment. Although science state partners voted on acceptance of final cut points, individual states had the option to adopt the consortium cut points or develop their own independent cut points.

Overview of DLM Science Assessment Design

Assessment Content

The DLM science assessment is based on Essential Elements (EEs) and linkage levels. The DLM EEs for science are specific statements of knowledge and skills linked to the grade-level expectations identified in the *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012; *Framework*) and the *Next Generation Science Standards, for States by States* (NGSS Lead States, 2013; NGSS). The purpose of the EEs is to build a bridge from those content standards to academic expectations for students with the most significant cognitive disabilities.

EEs for science consist of three linkage levels or access points to grade-level standards for students with the most significant cognitive disabilities. The linkage levels are Initial, Precursor, and Target. The Target linkage level aligns directly with the EE, while the other two linkage levels provide content at a reduced depth, breadth, or level of complexity. See the following example of science EE content at the three linkage levels.

Essential Element: EE.5-LS1-1 **Target Level:** Provide evidence that plants need air and water to grow.
Precursor Level: Provide evidence that plants grow.
Initial Level: Distinguish things that grow from things that don’t grow.

DLM science EEs are organized by science domain. Three domains are currently assessed: life science, physical science, and earth and space science. Each EE incorporates a topic and a scientific practice from the NGSS. In the above example EE, the topic is *organization for matter and energy flow in organisms* and the scientific practice is *engaging in argument from evidence*.

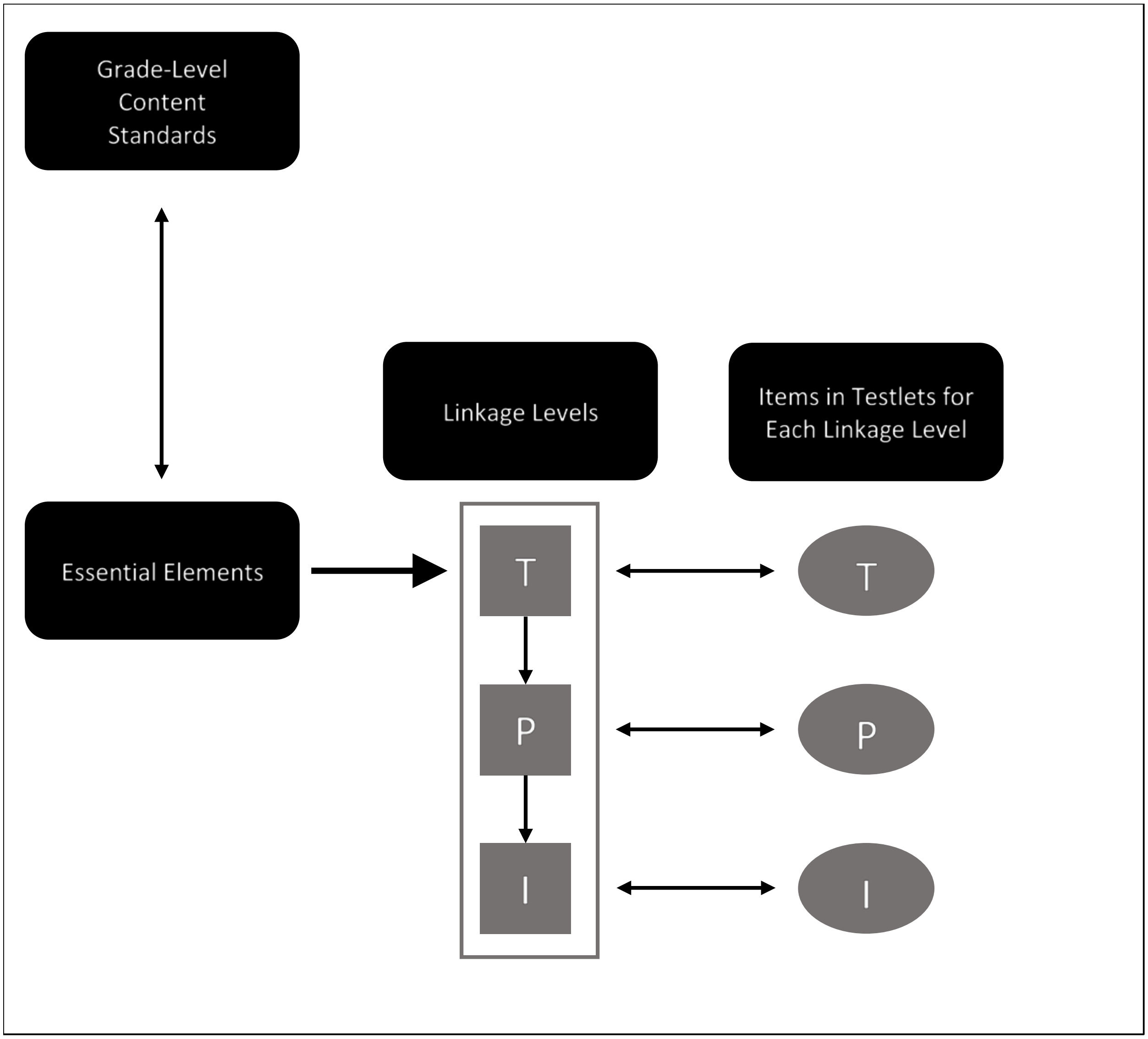
The science assessment system follows a year-end blueprint testing model, which has a consistent blueprint that is covered in its entirety in the spring testing window. Assessments are available in grade spans (3-5, 6-8, high school) and end-of-instruction (EOI) biology [1]. EEs were designed to be targets reached by the end of the grade span. Each science state requires assessment at different grade levels within the grade spans. As such, expectations for students in lower grades within a grade span could reasonably be lower than expectations for students at higher grades within the same span. Therefore, grade-specific achievement standards are needed. Based on TAC recommendation and a partner state vote, cut points were set at tested grade levels within the elementary and middle school grade spans (fourth, fifth, sixth, and eighth grades). In general, DLM science standard setting followed the same modified body of work methodology as was used in 2015 for the English language arts (ELA) and mathematics year-end and EOI models. For a detailed technical report on the methods used for the DLM ELA and mathematics standard setting process, please see *2015 Year-End Standard Setting: English Language Arts and Mathematics (Technical Report No. 15-03)*.

Assessment Design and Delivery

Each grade-level assessment is designed to assess a specific set of EEs. The EEs included in each blueprint can be found at http://dynamiclearningmaps.org/.

DLM assessments are delivered in testlets. Each testlet is comprised of items that align with a particular linkage level, as illustrated in Figure 2.

[1] States had the option of choosing which high school assessment to administer.


Grade-Level Content Standards
Linkage Levels
Items in Testlets for Each Linkage Level
Essential Elements
T
P
I
T
P
I

Note. T = Target; P = Precursor; I = Initial

Figure 2. Relationship between EEs, linkage levels, and items in testlets.

For the science assessment, the blueprint requires that all students be assessed on the same EEs. All students are assessed on testlets associated with the same EEs, but they are assigned testlets at different linkage levels so each student has an opportunity to independently demonstrate knowledge and skills. During the spring window, the linkage level of the student's first testlet was determined by the educator's responses to First Contact Survey items regarding the student's expressive communication skills. Each subsequent testlet linkage level was based on the student's performance on the previous testlet. If the student answered too few items correctly, the next testlet was at the next lowest linkage level. If the student answered all items correctly, the next testlet was at the next highest linkage level.

Scoring

Diagnostic Classification Modeling (DCM) is used to translate student responses to items into judgments about student mastery for each linkage level. For 2015-2016, students were considered masters of a linkage level if either: (1) their posterior probability from the DCM was greater than or equal to .80, or (2) the proportion of items that they answered correctly within the linkage level was greater than or equal to .80. Consistent with the ELA and mathematics scoring model, students who did not achieve mastery status for any tested linkage level were assigned mastery status for the linkage level that was two levels below the linkage level in which they were tested (unless the linkage level tested was either the Initial or Precursor levels, in which case, students were considered non-masters of all linkage levels within the EE). The scoring method for all content areas was discussed and approved by the DLM Technical Advisory Committee (TAC) during a conference call on July 21, 2015. [2]

Linkage level mastery status values were summed within and across EEs to obtain the total number of linkage levels mastered. Although the total number of mastered linkage levels is not a raw or scale score and should not be interpreted as an interval scale, the number of linkage levels mastered across EEs assessed was the metric translated into performance levels. Profiles used for standard setting were categorized by the number of linkage levels mastered across EEs. Further details on the development of profiles and the profile evaluation process are provided in subsequent sections.

Performance Levels and Policy Performance Level Descriptors

DLM science state partners chose to use the existing DLM performance levels and policy performance level descriptors (PLDs) originally developed for ELA and mathematics for science.

DLM state partners developed policy PLDs through a series of conversations and draft PLD reviews between July and December 2014. In July 2014, the state partners discussed general concepts that should be reflected in the PLDs and reviewed several examples of descriptors for three, four, and five performance levels. In fall 2014, the state partners indicated the number of levels they would require and gave feedback on additional iterations of PLDs that had been revised based on previous input. By December 2014, the PLDs were finalized. All states participating in the 2014-2015 operational assessment required four performance levels. The final version of policy PLDs are summarized in Table 2 below. The consortium-level definition of proficiency was At Target.

[2] More information about the psychometric model used for 2015-16 operational scoring is provided in Appendix A.

Table 2. Performance Level Descriptors.

Performance Level Descriptors
The student demonstrates ***emerging*** understanding of and ability to apply content knowledge and skills represented by the Essential Elements.
The student's understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements is ***approaching the target***.
The student's understanding of and ability to apply content knowledge and skills represented by the Essential Elements is ***at target***.
The student demonstrates ***advanced*** understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements.

Policy PLDs served as anchors for the standard-setting panelists during the panel process. This procedure is described in Chapter 2. The high-level process for developing grade- and content-specific PLDs is described in Chapter 6.

Organization of the Report

The remaining chapters of this report are organized into the following categories: methods, which includes a description of the overall approach and procedures; meeting preparation steps, which includes panelist recruitment and training; results, which includes panel-recommended cut points and associated impact data; statistical adjustment procedures and vertical articulation panel results; evaluations of panel recommendations; panelist evaluations of the meeting; and final recommended cut points.

Chapter 2: Standard Setting Methods

Rationale and General Approach

There is a history of selecting a standard-setting method based on the type of assessment. Because the DLM assessment is a unique alternate assessment system, the approach to standard setting was developed to be consistent with the DLM design while still relying on established methods, best practices recommended in the literature, and the *Standards for Educational and Psychological Testing* (2014).

There are several assessment design features that impacted the DLM standard-setting approach. A student-based standard-setting approach was judged to be more appropriate than an item-based approach for the following reasons:

- Modeling is used to support the order of linkage levels. Item difficulty statistics are not used to ensure correct ordering of content, so an item-based approach would not match the design of the test.
- DLM assessments are adaptive across testlets. Considering adaptive delivery and different forms for each EE/linkage level, it would be rare for students to receive completely identical testing experiences.
- A student-based approach supports the panelists' ability to make judgments about the student's mastery of the full range of skills rather than performance on a limited subset of items.
- The methods used for science are consistent with the methods used for other subject areas within the DLM assessment system.

For DLM assessments, the standard-setting approach leverages mastery classifications from the DCM model. The panel process draws from several established methods, including generalized holistic (Cizek & Bunch, 2006) and body of work (Kingston & Tiemann, 2012) but is unique to the DLM assessment. Other holistic approaches, such as the performance profile method (Perie & Thurlow, 2011), which takes into account the specific content mastered, would have been difficult to apply due to DLM partners' goal of reporting an overall performance level for each subject rather than subscores.

The DLM standard-setting approach relied on aggregation of dichotomous classifications of mastery of the knowledge and skills across EEs in the blueprint. This is different from assessments that use score scales, where standard setting involves identifying cut scores that are imposed on a theoretical, unidimensional continuum of knowledge in a subject.

Drawing from the generalized holistic and body of work methods, the DLM standard-setting process used a profile approach to classify student mastery into performance levels. Profiles provided a holistic view of student performance by summarizing mastery across the EEs and linkage levels. Cut points were determined by evaluating the total number of linkage levels mastered. Although the number of linkage levels mastered is not an interval scale, the process for identifying DLM cut points is roughly analogous to assigning a cut point along a scale score continuum.

Before making a final decision whether to use the profile approach, the DLM TAC reviewed a preliminary description of the proposed methods. At the TAC's suggestion, DLM staff conducted a mock panel process using this profile-based approach to evaluate the feasibility of the rating task and the likelihood of obtaining sound judgments using this method.

Although the DLM standard-setting approach is a unique hybrid of existing methods, the guidance in the *Standards for Educational and Psychological Testing* and recommended practices for developing, implementing, evaluating, and documenting the standard setting was followed (Cizek, 1996; Hambleton, Pitoniak, & Copella, 2012). For example, this report summarizes the rationale and procedures used to establish cut points (Standard 5.21), including evidence that the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way (Standard 5.23).

The following steps were used in the process and are described more fully in subsequent sections of this report.

1. Determine mastery and developing profiles
    a. Determine linkage level mastery
    b. Develop profiles of EE/linkage level mastery
2. Profile selection
3. Panelist profile ratings
4. Statistical analysis of panelist ratings
5. Review of impact data
6. Adjust cut points for cross-grade consistency

Determining Mastery and Developing Profiles

Because of the unique nature of the DLM assessment, student performance on linkage levels within EEs must be aggregated to create profiles of student learning. There were two steps in the standard-setting process to determine student performance level. First, student mastery at the linkage level was determined for each EE using the DCM approach. Then, profiles of student linkage level mastery were used during the standard-setting process to set cut scores. The first step addressed mastery thresholds that can be applied wholesale, and the second step set performance level cut points using the panel process. The threshold criteria were originally developed for the DLM ELA and mathematics assessments. To be consistent, science applied the same criteria wherever relevant. Descriptions of the criteria used in each step for science are provided in the following sections. For further detail on the rationale for the criteria used for ELA and mathematics, please see Karvonen, Clark & Nash (2015).

Define Mastery at the Linkage Level

Mastery classifications were derived from applying an agreed-upon criterion to students' posterior probabilities from the DCM calibration. These posterior probabilities are referred to as linkage level mastery probabilities.

A student's classification as master or non-master was made with a level of certainty that was reflected in the posterior probability. The acceptable level of certainty (i.e., the posterior probability) had to be identified before additional judgments could be made in the standard-setting process. Maximum uncertainty occurs when the probability is .50, and maximum certainty when the probability approaches 0 or 1. Considering the risk of false positives and negatives, the threshold used to determine mastery classification was .80. That is, students with linkage level mastery probabilities ≥ .80 were considered masters of the level while students with probabilities < .80 were considered non-masters of the level.

For each linkage level, a mastery status of 0 or 1 was obtained based on the student's probability of linkage level mastery. Using.80 as the cutoff for linkage level mastery, all students with a probability greater than or equal to.80 received a linkage level mastery status of 1, or mastered. All students with a probability lower than .80 received a linkage level mastery status of 0, or not mastered. Students were also considered masters of a linkage level if the proportion of items that they answered correctly within the linkage level was greater than or equal to .80. If the student tested but did not master a linkage level, then mastery status would be assigned at two levels down from the tested level.

Develop Profiles of EE/Linkage Level Mastery

The threshold values from step one were applied to create profiles of student mastery, which summarize linkage level mastery by EE. Profiles were created using data for each grade span. Each profile listed all the EEs from the blueprint containing between nine and ten EEs. The three levels of mastery were included as columns on the profile, ranging from the Initial linkage level up to the Target linkage level. Green shading indicated that a linkage level was mastered (the threshold was met) for students matching that profile. Blue shading indicated that the EE was tested but no linkage level was mastered, and no shading indicated that the EE was not tested.

Appendix B: Sample Profile Based on Judgments about Linkage Levels: Middle School Science provides an example of a science profile for a middle school student. The profile demonstrates one example of the possible skills mastered for a student who has mastered 14 linkage levels, as evidenced by the green shading in 14 boxes.

The maximum linkage level mastery value was determined by the blueprint: the count of EEs times three linkage levels was considered the highest total linkage level value. All grade spans in science have nine EEs, with 27 as the highest total linkage level value, with the exception of end-of-instruction biology, which has 10 EEs, for a total of 30 possible linkage levels.

Profile Selection

In order to select exemplar profiles for inclusion in standard setting, a program was written in *R* to determine the highest linkage level the student mastered for each EE and sum them to get the total linkage level mastery value. As a quality control effort, psychometric staff members ensured that the results of the program were expected based on the input data. Specifically, checks were conducted to determine that the program accurately applied mastery thresholds and correctly determined the highest linkage level mastered by the student.

Profiles were available for all students who participated in the spring window by May 12, 2016 ($N = 20{,}448$, $n_{3\text{-}5} = 5{,}455$, $n_{6\text{-}8} = 5{,}622$, $n_{9\text{-}12} = 5{,}098$, $n_{\text{Bio}} = 1{,}312$). A program was written in *R* to identify the frequency with which each precise profile (i.e., pattern of linkage level mastery) occurred in this population. Based on these results, the three most common profiles were selected for each possible total linkage level mastery value (i.e., total number of linkage levels mastered) for each grade span. For example, the program identified the three most common ways to have mastered 18 linkage levels for the elementary grade span. To ensure that the exemplar profiles were not overly similar, the program identified profiles where different linkage levels were mastered for at least three EEs.

In instances where data was not available at a specific linkage level value, (e.g. no students mastered exactly 26 linkage levels for a grade and content area), profiles were based on simulated data. The science content team used adjacent profiles for reference and created simulated profiles that represented likely patterns of mastery. This approach was consistent with the process used for ELA and mathematics standard setting in 2015. Fewer than 4% of all the profiles developed were simulated. Simulated profiles were not distinguishable from those based on real student data.

Profile Rating Procedures

Exemplar profiles of student mastery were compiled in folders for panelist ratings. Two types of folders were prepared for standard setting: range-finding folders and pinpointing folders. After panelists familiarized themselves with performance levels during training, the range-finding process followed. The purpose of range finding was for panelists to assign general divisions between performance levels after reviewing a limited set of profiles from points along the distribution of total linkage levels mastered. These samples were comprised of profiles at intervals of five linkage levels mastered (e.g., a total of 5, 10, 15, and 20 linkage levels mastered). After panelists determined the ranges of linkage levels mastered where cut points were likely to be found, panelists completed the pinpointing process. The purpose of pinpointing was for panelists to evaluate the additional exemplar profiles with the goal of identifying specific cut points. Profiles for seven adjacent levels within the range determined from the range-finding process were distributed to the panelists for pinpointing. For both the range-finding and pinpointing phases, panelists completed an independent round of ratings, reviewed their results and discussed them, then completed a second round of independent ratings. The results at the end of the second pinpointing round served as the group-recommended cut points. Finally, panelists independently evaluated the group-recommended cut points and indicated their level of confidence with each cut point. Additional detail about these procedures is provided in the Procedures section of Chapter 4.

Statistical Analysis of Panelist Ratings

Both the range-finding and pinpointing exercises utilized logistic regression analyses to identify appropriate ranges and calculate cut points, respectively. Logistic regression models the relationship between an independent variable, number of linkage levels mastered in this case, and the probability of being classified into a category, such as performance level **approaching** or above.

The primary goal of using logistic regression as the analytical procedure is to identify the number of linkage levels mastered where the likelihood of being assigned to a given performance level equals or exceeds that of being assigned to the next performance level or higher (where $p = .50$). In other words, rather than determining the likelihood of resulting in a specific level, given a number of linkage levels mastered, the goal was to find the likelihood of being assigned to a level **or higher**, given a number of linkage levels mastered. For example, if logistic regression indicated that the likelihood of panelists assigning a profile with 20 linkage levels mastered to performance level **approaching** or higher is 7 out of 12 (about 58%), it could be concluded that 20 linkage levels mastered would be a good cut point to differentiate **emerging** from **approaching**.

For pinpointing, the range of profiles was calculated by taking the value determined during range-finding plus and minus three for a total of seven different profiles each representing a different number of linkage levels mastered. Using this narrowed but more informative range of exemplar profiles, logistic regression was again used during the pinpointing process to determine the point at which the probability of being assigned to each performance category or higher was .50. The predicted values from this process were used as the recommended cut points for each level.

In some cases, the logistic regression analysis did not yield a useful result. Because this analysis largely depends on identifying areas of maximum disagreement between panelists across two performance categories to identify the point at which the probability is .50, logistic regression failed for any case where all of the panelists within a group had unanimous agreement on profile ratings. In these cases, on-site psychometricians reviewed the panelist group ratings and visually identified where the obvious inflexion point occurred. The value where the shift in ratings moved from one category to the adjacent category was used as the recommended group cut point.

The regression analyses to obtain the cut points were carried out in Excel using the same facilitator workbooks in which the original data were tallied and transformed to logistic functions. The facilitator workbooks are discussed in more detail in the Procedures section of Chapter 4.

The panelists' independent evaluations of the group-recommended cut points were summarized and evaluated using descriptive statistics. The purpose of evaluating the independent ratings was to identify any places where the median independent recommended cut points differed from the group-recommended cut points.

Impact Data

Impact data was calculated by grade based on total number of linkage levels mastered. The percent of students who would be classified at each performance level based on the panelists' recommended cut points was calculated and presented to the panelists at the conclusion of the final pinpointing ratings. No further discussion was held at that time; rather, a subsequent step was conducted, in which a cross-grade panel reviewed and discussed impact data patterns across all grade levels (discussed in the next section).

State partners served as the policy group for reviewing impact data. The state partners, who are members of the DLM science consortium governing board, have varying roles within the special education and assessment departments in their state education agencies. These partners were not only knowledgeable of the DLM assessment system, but also of their own states' educational policies and student populations. State partners discussed recommended cut points and impact data with their internal stakeholders and reviewed input from the DLM TAC before participating in consortium-level discussions. Additional details regarding recommended cut points, impact data, and cut point adjustments are provided in Chapter 5.

Vertical Articulation Panel

Once the panel-recommended cut points were set, two representatives from each panel (except end-of-instruction biology [3]) convened to conduct a cross-panel review and

[3] End-of-instruction biology was not included in the vertical articulation process, as it was not expected that students in one course were representative of the students in the general high school grade span and there was no reason to expect that a single EOI biology assessment was somehow contiguous to a previous grade-level, multi-domain assessment.

discussion of the panel-recommended cut points, statistically adjusted cut points (methodology discussed in a subsequent section), and the associated impact data for each. The process began with a discussion of panelists' content-based rationales for their ratings and their panel's recommended cut points across grade levels. Next, panel-recommended cut points and statistically adjusted cut points (procedures for adjustment are described in Chapter 5) with impact data for each were presented for all grade-level panels and high school. After a whole group discussion about the system of cut points focusing on content-based rationales for results, the panel's conclusions and final recommendation were documented.

Evaluation Procedures

The standard-setting procedures were evaluated using procedural, internal, and external criteria as described by Hambleton & Pitoniak (2006). Each category contains several sub-categories. Relevant sub-categories are addressed individually.

Procedural Criteria

Explicitness. The standard-setting process was explicitly defined prior to the standard-setting event. Facilitators used a guide with detailed instructions for each step in the process. As part of the training for the event, all facilitators went through a mock standard setting where they used the intended process to ensure that there was an understanding of how the process should occur.

Practicability. To evaluate the use of the intended standard-setting approach, a mock panel convened to test the process and evaluate its ease of use and likelihood of generating the intended results. In instances where the outlined procedures were inadequate (e.g., the logistic regression failed due to unanimous panelists recommendations), solutions were quickly implemented without creating confusion for the facilitators or panelists.

Implementation of Procedures. The selection of panelists was completed in the most objective way possible while also ensuring adequate coverage of content areas and grade levels. During the panel meeting, staff used a step-by-step guide to ensure fidelity of implementation. Where procedures had been revised since the ELA/Math standard setting that was conducted in 2015, staff and panelists were trained on the revisions. Additionally, DLM staff members who were not facilitating specific panels observed the standard-setting event to verify that the specified procedures were being implemented correctly. Panelist selection and assignment is described in Chapter 3. The training of the panelists is detailed in Chapter 4.

*Panelist Feedback***.** After receiving training for the standard-setting event, nearly all panelists reported "Good" or "Excellent" understanding of important and relevant ideas. This included the purpose of standard setting, how DLM assessments assess content knowledge, and how scores are calculated and reported. Notably, no panelists reported "Poor" understanding for any of the key ideas assessed. Further details are presented in Chapter 4.

Documentation. When developing this standard-setting method, documentation was kept on the proposed techniques, associated rationales, and TAC and state feedback. Documentation was also kept on all stages of the process, including panelist recruitment and selection, training, and implementation. This technical report is largely based on source documentation.

Internal Criteria

Consistency Within Method. The variability of panelists' final pinpointing ratings and their final independent ratings were reported. Standard errors are presented in Chapter 5.

Interpanelist Consistency. Due to the nature of the standard-setting method used (i.e., logistic regression to identify areas of maximum disagreement as potential cut points), interpanelist consistency was not the desired outcome. However, there was an expectation that panelists would converge towards an increasingly narrow range of profiles to identify the cut point. Evidence of convergence is described in Chapter 5.

External Criteria

Reasonableness of Performance Levels. The panel-recommended and adjusted cut points, with the corresponding impact data, were presented to state partners to ensure their reasonableness. Further details of this process may be found in Chapter 5.

Reasonableness of Standard-Setting Process. The proposed standard-setting process was presented to the TAC prior to the event to ensure its reasonableness, and a TAC member attended the standard-setting event to ensure its fidelity to the proposed process.

Chapter 3: Standard Setting Panel Meeting Preparation

Panelist Recruitment

DLM staff drafted and distributed a recruitment letter to participating DLM states in March 2016. The recruitment letter is included in Appendix C: Standard Setting Panelist Recruitment Letter and Survey. Participating states for standard setting included those that were operational in 2015-2016. States were responsible for distributing the letter within their state to recruit potential panelists. Some states elected to distribute the list narrowly to constrain the number of potential panelists to only those they recommended. Others distributed the call more broadly within the states.

DLM staff sought panelists with content knowledge and expertise in the education and outcomes of students with significant cognitive disabilities, including educators as well as school and district administrators. Other subject matter experts, such as higher education institution faculty or state/regional educational staff, were also suggested for consideration.

All potential panelists were asked to complete a survey. Survey items included basic demographic information as well as areas of expertise and years of experience. In addition, volunteer panelists were asked to indicate whether they were willing to commit to advance training (up to four hours during the first two weeks in June) and whether they would be available to attend the on-site meeting from June 15– 17, 2016. See the survey in Appendix C: Standard Setting Panelist Recruitment Letter and Survey.

Selection of Panel Participants

DLM staff received 164 total responses to the survey. All survey responses were evaluated in April 2016 to assign volunteers to panels. Panelists' home state; diversity of experience in education; and levels of expertise with science content, education, and students with severe cognitive disabilities were given priority in the selection of panelists. Race/ethnicity, gender, and urbanicity were also considered.

Forming Panels

Six panels were created from the pool of volunteers, with representation as spread across the states as possible. Specifically, a panel was created for each of the following grades, grade span, and course: 4, 5, 6, 8, high school (9–12), and biology.

Each panel (with the exception of high school and biology) consisted of four panelists that had teaching experience and expertise at their assigned grade level or grade span. The high school panel consisted of eight panelists. The end-of-instruction biology panel consisted of eight panelists from Oklahoma, since it is the only consortium member state that participated in the end-of-instruction biology assessment.

Panelist Characteristics

The 32 panelists who participated in standard setting represented varying backgrounds, as summarized in Table 3. Most of the selected panelists were classroom educators. Panelists had an average of 16.2 years of experience in the field of education and had a range of years of experience with science content and working with students with significant cognitive disabilities. The maximum, minimum, and mean years of experience are presented in Table 4. The number of panelists who taught or worked with students in each disability category are displayed in
Table 5.

Table 3. Panelist Demographic Characteristics

Demographic Characteristics	*n*
Gender	
Female	29
Male	3
Race	
African American	3
American Indian/Alaska Native	3
Asian	2
Hispanic/Latino	2
Native Hawaiian/Pacific Islander	1
White	21
Professional Role	
Classroom Teacher	23
Building Administrator	0
District Staff	6
State Education Agency Staff	2
University Faculty/Staff	2
Other	8
Total	32

Table 4. Panelist Years of Experience

Experience Type	*M*	Min	Max
Students with Significant Cognitive Disabilities	14.3	2.0	30.0
Science	13.2	1.0	30.0

Table 5. Number of Panelists Who Taught Students in each Disability Category

Disability	Count
Blind/Low Vision	22
Deaf/Hard of Hearing	20
Emotional Disability	26
Mild Cognitive Disability	28
Multiple Disabilities	30
Orthopedic Impairment	24
Other Health Impairment	28
Severe Cognitive Disability	30
Specific Learning Disability	25
Speech Impairment	29
Traumatic Brain Injury	24

Note: More than one disability category could be selected.

Nearly half of the participants had experience with setting standards for other assessments (15). Some panelists already had experience with DLM, either from writing items (8) or externally reviewing items and testlets (10). Only one panelist reported having less than one year or no experience with alternate assessments; that panelist was university faculty/staff with 19 years of experience with science content.

Panel Facilitator Training

All staff, including facilitators, room leads, and supporting staff, participated in a one-hour orientation meeting regarding the purposes and outcomes of standard setting. Staff reviewed a high-level overview of the procedure. Following orientation, facilitators read a description of the training range-finding and pinpointing procedures. During the next training session, panel facilitators received a detailed agenda and scripts to be used for the standard-setting process. Five of the six facilitators had previously served as a facilitator during the 2015 standard-setting event for ELA and mathematics. The new facilitator had previous experience with standard settings that followed similar procedures, as well as the 2015 mock run-through of the standard-setting process. All facilitators practiced leading a group using the agenda and scripts and learned how to enter panelist ratings in the facilitator workbook. The agenda and scripts were adjusted prior to the standard-setting panel meeting based on this run-through. Debriefs were also held each day of the panel meeting to review any remaining questions.

Chapter 4: Standard Setting Panel Meeting Procedures

Panelist Training

Advance Panelist Training

All panelists participated in a training module in advance of the standard setting meeting. The purpose of this training was to give panelists a general overview of the DLM assessment system ahead of time so that on-site training could focus on the panelists' specific grade/content area assignment and panel procedures. After introducing the purpose of standard setting and expectations for confidentiality, the advance training addressed the following topics:

1. Students who take DLM assessments
2. Content of the assessment system, including EEs for science, domains and topics, linkage levels, and alignment
3. Accessibility by design, including the framework for the DLM assessment's cognitive taxonomy and strategies for maximizing accessibility of the content; the use of the Personal Needs and Preferences (PNP) profile to provide accessibility supports during the assessment; and the use of First Contact Survey to determine linkage level assignment
4. Assessment design, including item types, testlet design, and sample items from various linkage levels in science
5. An overview of the assessment model, including test blueprints and the timing and selection of testlets administered
6. A high-level introduction to two topics that would be covered in more detail during on-site training: the DLM approach to scoring and reporting and the steps in the standard setting process.

The advance training was available online, on demand during the ten days prior to the standard-setting meeting. All panelists completed the required training before arriving for the on-site panel meeting.

After viewing the training videos, panelists completed a survey where they rated their understanding of key topics. The results are summarized in Table 6. Panelists reported feeling most comfortable with areas referencing the characteristics of students taking DLM assessments, the expectations for maintaining security of information during the training, and standard setting. Since most panelists were also educators who administered DLM assessments, these were likely areas where they had direct experience. Panelists reported being less comfortable with the more technical aspects of how testlets measured content and calculation and reporting of results.

Table 6. Panelist Self-Assessments after Completing Advance Training

Understanding of:	Poor	Fair	Good	Excellent
Characteristics of students who take DLM assessments	0	0	7	29
The purpose of standard setting	0	1	13	22
Essential Elements and linkage levels	0	2	13	21
Expectations for maintaining security of information during training and standard setting	0	0	1	35
How testlets measure the intended content	0	2	15	19
How testlets are made accessible to students from across the DLM population	0	0	11	25
What a student is expected to do during a DLM assessment	0	0	10	26
How results are calculated and reported	0	1	23	12

Panelists also rated their overall preparation for the next phase of training and whether their understanding was sufficient to make judgments about student results. All panelists ranked themselves as either **very prepared** (23) or **somewhat prepared** (13) for the next training at standard setting, and 100% of panelists believed their knowledge to be sufficient to make judgments about student performance and assessment results.

On-Site Panelist Training

Additional panelist training was conducted onsite. The purposes of on-site training were twofold: (1) to review advance training concepts that panelists had indicated less comfort with, and (2) to prepare panelists for their responsibilities during the panel meeting. Since the majority of panelists indicated a high degree of comfort with advance training concepts, the first part of on-site training was a high-level review of expectations for confidentiality and test security, the organization of academic content, and testlet design. Prior to training on the standard-setting procedures, panelists were prompted to ask questions about any of the topics from the advance training.

Training on the standard-setting panel procedures included the following topics:

1. How results are calculated and displayed in mastery profiles for standard setting, including guidance about appropriate interpretations of the contents of mastery profiles
2. An overview of the standard-setting process including the policy PLDs, terms used during the standard-setting process, the key question panelists would ask themselves when completing ratings, and the range-finding and pinpointing procedures
3. An overview of the event's activities, from training to final evaluation
4. Roles and responsibilities of everyone present for the panel meeting
5. Discussion of the contents and use of the policy PLDs
6. Presentation of the resource materials panelists should refer to when familiarizing themselves with mastery profiles

After the large group presentation on these topics, the trainer introduced the practice activity to be completed at each panel table. The training activity consisted of range finding using training profiles for just a few total linkage levels mastered (e.g., 5, 10, 15, 20). Each table trained using sample profiles for the grade/course for which the panelists would be setting standards. Table facilitators walked panelists through the process of using their resource materials to familiarize themselves with the EEs and linkage levels for that grade/course. Once panelists were ready, the facilitator then introduced the contents of the training folder (i.e., the training profiles and rating forms) and reminded panelists how to complete the rating form. Once all panelists completed the practice activity, they had opportunities to debrief at the table. Two smaller group discussions were also conducted (based on timing of completion of the practice activity) to discuss the process and provide guidance on expected patterns of ratings across ranges of profiles.

Since all panels were expected to work on range finding during the first day, more in-depth training on the pinpointing procedure was reserved for the second day. Training on the second day also covered procedures for capturing information to be used for grade-specific PLDs.

Additional detail about on-site training is provided in the agenda and training slides in Appendix D: Panel Training and Materials.

Materials

Panelist Resources

Each panelist received a resource notebook with materials to use in training and during the rating process. The resource notebook contained

- a standard-setting flowchart,

- an annotated sample mastery profile,
- a PLD handout,
- hints for making ratings,
- instructions for completing rating forms,
- diagrams of the elements of the DLM system, and
- a glossary of DLM and standard-setting terms.

When familiarizing themselves with each grade's EEs and linkage levels, panelists also used the following resources:

- EE tables that outlined each EE's associated state standard for general education (using the NGSS coding system), connections to science practices, crosscutting concepts as well as connections to DLM ELA and mathematics EEs
- The science and engineering practices (adapted from the Next Generation Science Standards; Achieve, 2013) that are embedded in the DLM science EEs
- A blank mastery profile for that grade (i.e., one that contained EEs and linkage level descriptions but no mastery shading)
- The blueprint for that grade

Panelists also had access to sample testlets for any EE/linkage level assessed in a grade. Upon request, facilitators displayed sample testlets in the online content management system.

Training Materials

Training folders were prepared with exemplar profiles of student mastery for grade-specific panels. The training folders included six exemplar profiles: two profiles with 7 levels mastered, two profiles with 14 levels mastered, and two profiles with 21 levels mastered. Two examples were included at each linkage level mastery amount to show how students with the same number of linkage levels might achieve that number by mastering different EEs or linkage levels. The training folders also contained sample rating sheets.

Range-Finding Materials

Range-finding folders were prepared with exemplar mastery profiles from across the range of student performances for the specific grade being reviewed. The number of profiles varied depending on the number of linkage levels on the blueprint. All grade spans, with the exception of the end-of-instruction biology blueprint, have nine EEs and 27 linkage levels; biology has 10 EEs and 30 linkage levels. Exemplar profiles were provided in five-number increments. For example, in a grade with nine EEs and therefore 27 linkage levels, the range-finding folder included profiles for students who mastered 5, 10, 15, 20, and 25 linkage levels.

Profiles were ordered in the folder according to the total number of linkage levels the student mastered. There were three exemplar profiles for each available level of mastery. In the previous example for a grade with 27 possible linkage levels, a total of 15 profiles would be included in the folder spanning the five possible linkage level values included.

All exemplar profiles were numbered to ease discussion.

Pinpointing Materials

The pinpointing folders contained profile exemplars for a reduced range of levels around potential cut points. For each cut point, exemplar profiles were included at seven levels, including the number closest to the suggested cut point determined in range finding and three above and below that number. For example, if range finding identified that a given cut point should be somewhere around 20 linkage levels mastered, the folder would contain profiles with 17, 18, 19, 20, 21, 22, and 23 linkage levels mastered. A folder contained three profiles for each number of linkage levels mastered (i.e., multiple ways students have actually demonstrated the same number of linkage levels mastered), for a total of 21 profiles at the seven levels. Any profiles that were used in range finding were reused in pinpointing (e.g. the three profiles reviewed for 20 linkage levels mastered during range finding were also included in the pinpointing folder).

Rating Forms

Rating forms for each of the range-finding and pinpointing processes were provided in the panelists' folders. One range-finding rating form and one pinpointing rating form were provided for each subject and grade-level set of cut points. Each form contained columns for round one (first) and round two (final) ratings. Example range-finding and pinpointing rating forms are provided in Appendix E: Example Rating Forms for Range Finding and Pinpointing.

Evaluation Form

An evaluation form was provided to panelists for the purpose of obtaining panelists' independent evaluations of group recommended cut points and panelists' evaluations of the overall standard-setting training and meeting. The evaluation was provided to panelists on the closing day of the standard-setting meeting and is provided in Appendix F: Panelist Meeting Evaluation Form.

Procedures

Both the range-finding and pinpointing procedures consisted of two rounds of ratings. Panelists reviewed the exemplar profiles, independently rated each profile for round one ratings, discussed ratings as a group, and then independently rated each profile again for round two ratings. Throughout both range finding and pinpointing, panelists were instructed to use their best professional judgment and consider all students with significant cognitive disabilities to determine which performance level best described each profile.

Details of the final procedures used for determining cut points is provided in the subsequent sections.

Range Finding

During the range-finding process, panelists reviewed a limited set of profiles to assign general divisions between the performance levels. The goal of range finding was to locate ranges (in terms of number of linkage levels mastered) where panelists agreed that approximate cut points should exist.

These are the procedures the panelists followed for range finding.

1. Panelists independently evaluated the profiles in the range-finding folder and identified the performance level that best described each profile. They recorded their decision for each exemplar profile on their rating sheet.
2. Once all panelists completed their ratings, the facilitator obtained the performance level recommendations for each profile by a raise of hands. The facilitator recorded the counts in the facilitator workbook, which was projected for the group to view. One panelist at each table was assigned to check that the values were entered correctly to ensure accurate data entry.
3. After table discussion of how panelists chose their ratings, the panelists were given the opportunity to adjust their independent ratings if they chose. A second round of ratings were recorded and shared with the group. Again, the facilitator entered values in the facilitator workbook, and the designated panelist confirmed their accuracy.
4. Using the round two ratings, built-in logistic regression functions calculated the probability of a profile being categorized in each performance level conditional on number of linkage levels mastered, and the most likely cut points for each performance level were identified.
5. Psychometricians reviewed every workbook before the group began the pinpointing process to ensure no errors were present and to check that the logistic regression had successfully determined a reasonably appropriate approximate cut point. In instances where the logistic regression function could not identify a value (e.g. the group unanimously agreed on the categorization of profiles to

performance levels), psychometricians evaluated the results to determine the approximate cut point based on the panelist recommendations.

Pinpointing

During pinpointing, panelists reviewed additional profiles to refine the cut points. The goal of pinpointing was to pare down to specific cut points in terms of number of linkage levels mastered within the general ranges determined in range finding, not relying on conjunctive or compensatory judgments.

These are the procedures the panelists followed for pinpointing.

1. Folders containing the profiles for the seven levels, including and around the cut point value identified during range finding were distributed to the panelists.
2. Panelists independently evaluated the profiles in each folder and assigned each a performance level—those in the higher level and those in the lower level. Panelists entered their recommendations on their pinpointing rating sheet.
3. Once all panelists completed their ratings, the facilitator obtained the recommendations for each profile by a raise of hands. These counts were entered into the projected facilitator Excel sheet. The identified panelist checker confirmed all values were entered correctly.
4. After discussion of the ratings, a second round of rating commenced. Panelists were given the opportunity to adjust their independent ratings if they chose.
5. The facilitator collected final ratings by show of hands. The panelist checker confirmed values were entered correctly.
6. Using the second round's ratings, built-in logistic regression functions calculated the probability of a profile being categorized in each performance level conditional on number of linkage levels mastered, and the most likely cut points for each performance level were identified.
7. Psychometricians reviewed every workbook at the close of the pinpointing process to ensure values were obtained accurately. In instances where the logistic regression function could not identify a value (e.g. the group unanimously agreed on the categorization of profiles to performance levels), psychometricians evaluated the results to determine the final recommended cut point based on the panelist recommendations.

Chapter 5: Results

This chapter summarizes the panel-recommended cut points, evaluation evidence regarding the panel process, impact data, and the final results.

Panel-Recommended Cut Points and Associated Impact Data

Table 7 includes a summary of the cut point recommendations reached by the panelists following the range-finding and pinpointing process. Note that the last column represents the maximum number of linkage levels that are possible based on blueprint requirements for each grade.

Table 7. Panel Recommended Science Cut Points

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Maximum Number of Linkage Levels
4	9	16	22	27
5	11	18	25	27
6	9	15	22	27
8	11	16	23	27
9-12	9	17	24	27
Biology	9	15	22	30

Impact data was calculated using the linkage level mastery status and total number of linkage levels mastered on each tested EE for all students. Duplicate student records, which could have occurred based on school or district data management practices, were removed using the following rule:

> *Remove duplicates when the following fields were all identical across rows: student ID, state, grade level, and number of linkage levels mastered.*

This step prevented the same student's linkage level mastery status from being used multiple times in the calculation of the impact data. This means that if a student was rostered to multiple educators, the data were only included once. Students who were rostered in the system but did not test on any EEs were not excluded from the data file. However, because these students had no scores, their inclusion did not influence the frequency distributions of the impact data. Once duplicate records were removed, the frequency distributions of students at each performance level were calculated for grade level.

Table 8 displays the frequency distributions associated with the panel-recommended cut points. The majority of students were categorized as either Emerging or Approaching the Target performance levels with the exception of end-of-instruction biology, where there was a more even distribution across the four performance levels. The distribution of students observed in biology is was consistent with those in DLM ELA and mathematics end-of-instruction courses. The limited number of states participating in end-of-instruction courses (i.e., one state in science) may have contributed to a lack of representation of the student population. As noted previously, panelists were presented the impact data after their final pinpointing ratings were complete but no further discussion was conducted at that time.

Table 8. Percentages of Students in Each Performance Level Based On Panel Recommended Cut Points

Grade	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target/Adv (%)
4	59.4	27.0	10.6	2.9	13.5
5	62.9	20.3	12.5	4.2	16.7
6	45.4	30.6	21.0	3.0	24.0
8	57.7	20.4	18.7	3.2	21.9
9-12	59.6	26.6	12.0	1.8	13.8
Biology	32.3	20.0	22.3	25.5	47.8

Convergence

The purpose of range finding and pinpointing was to identify the specific number of linkage levels mastered that would differentiate student performance into each of the four performance levels. Through each round of discussion and ratings, panelists narrowed in on the range in which the cut point could be identified. Due to the nature of the statistical analysis method used, inter-panelist consistency was not the desired outcome for a single round; however, there was an expectation that panelists would converge toward an increasingly narrow range of profiles to identify the cut point. To illustrate the degree to which panelists converged upon an agreed upon cut point, box and whisker plots are displayed in Appendix G: Convergence Plots for Range-Finding and Pinpointing Ratings. These plots convey the median, first and third quartiles, and range of the frequencies with which each number of linkage levels mastered was classified into each of the four performance levels.

Overall, the plots support the claim that the panel process worked as intended. In general, the ranges of profiles categorized into each performance level narrowed from round one to round two during both range finding and pinpointing.

Standard Errors of Pinpointing Ratings

Following the standard-setting event, standard errors were computed to evaluate the results. This method was based on the frequency distributions of panelists' final pinpointing ratings and was accomplished by dividing the standard deviation of the frequencies of panelists' final pinpointing ratings by the square root of the number of total ratings. Table 9 displays the standard errors for the distribution of final pinpointing ratings.

Table 9. Standard Errors for Science Final Pinpointing Ratings

	G4	G5	G6	G8	G9-12	Biology
Emerging	0.184	0.174	0.000	0.204	0.115	0.140
Approaching	0.330	0.228	0.215	0.217	0.191	0.162
Target	0.202	0.210	0.215	0.267	0.157	0.161
Advanced	0.163	0.104	0.184	0.162	0.073	0.109

Statistical Adjustment

Procedure

Despite evaluative evidence that was generally supportive of the panel-recommended cut points, these recommendations are based on the work of single panels. Each panel is a sample of possible experts. In theory, some variability in recommended cut points would be expected with a different sample, and each sample's recommendation would be an estimate of the true cut point.

To mitigate the effect of sampling error and issues related to a system of cut points across a series of grade levels, many testing programs consider impact data in the grade at question *and* contiguous grades. The logic is that under most circumstances (especially when there is no significant shift in demographics), students in bordering grades should have similar distributions within performance levels. Dramatically different distributions are likely due to sampling error and not differences in true cut points.

While the DLM science assessments were designed and administered at three grade spans (elementary, middle school, and high school) and one end-of-instruction biology assessment, standards were set for grade-specific panels for grades 4, 5, 6, and 8. Statistical adjustments were made to the grade-specific panel-recommended cut points in an effort to systematically smooth distributions within the system of cut points being considered. No adjustments were made for EOI since there was no reason to expect that the students taking biology were in any way representative of the students in the general high school grade span. Similarly, there was no reason to expect that a single EOI biology assessment was contiguous to a previous grade level, multi-domain assessment.

The following steps were applied to each grade level.

1. Create a frequency distribution of the number of linkage levels mastered (from low to high). The number of possible linkage levels is 27 for each grade.

2. Calculate cumulative proportions from low to high.
3. Perform a probit transformation (*z*-score associated with the cumulative proportion of students) for each number of linkage levels mastered. Because at the top of the distribution (proportion equal to 1) a finite *z*-score cannot be calculated, to perform subsequent calculations, top *z*-scores were defaulted to 3.5.
4. Find the *z*-score associated with the raw cut point of interest (for example, Approaching/Target).
5. Create a weighted rolling average of z-scores for the cut point of interest using a weight of 0.5 for the grade of interest and 0.25 for contiguous grades.
$$\sum w_i Z_i \Big/ \sum w_i$$
At the ends (grades 4 and high school) there cannot be a symmetric set of three grade levels involved in the rolling average.
6. Using the table of probit-transformed cumulative proportions, look up the raw number of linkage levels mastered for which the *z*-score is closest to the weighted rolling average of *z*-scores. The closest *z*-score was selected instead of the lowest *z*-score to prevent systematically decreasing the proportion of students in the higher category over the system of cut points.

Adjusted Cut Points and Associated Impact Data

Table 10 and Table 11 summarize the adjusted cut points that used the methods described above and the impact data for those adjusted cut points. Frequency distributions for the impact data of the adjusted cut points were calculated using the same process as described for the panel-recommended cut points.

The approach used did decrease the between-grade variability as expected. All but one adjustment lowered the cut point by one point. The sixth grade cut point between Emerging and Approaching was the only cut point that increased one point as a result of the statistical adjustment.

Table 10. Statistically Adjusted Science Cut Point Recommendations

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Maximum Number of Linkage Levels
4	9	15	21	27
5	10	17	25	27
6	10	15	21	27
8	10	16	23	27
9-12	8	16	23	27

Note. Cut points for biology were not statistically adjusted.

Table 11. Percentages of Students in Each Performance Level Based on Adjusted Cut Point Recommendations

Grade	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target/Adv (%)
4	59.4	24.0	12.6	4.0	16.6
5	58.5	21.9	15.4	4.2	19.5
6	51.7	24.4	19.2	4.8	24.0
8	52.6	25.5	18.7	3.2	21.9
9-12	54.1	29.6	13.0	3.3	16.3

Note. Cut points for biology were not statistically adjusted.

Vertical Articulation Panel Process

The vertical articulation panel was comprised of representatives from each panel (except end-of-instruction biology) who were tasked with evaluating both the panel recommended and statistically adjusted sets of cut points and associated impact data. In reviewing and considering the cut points and impact data across all grade levels and thinking about how skills are taught from one grade to the next, the vertical articulation panel made a strong cross-grade content-based rationale for recommending all of the adjusted cut points, with the exception of one cut point. Specifically, they recommended retaining the panel recommended cut point for the sixth grade cut between Emerging and Approaching the Target. As the adjusted cut points at this level for sixth and eighth grades were the same, they chose to retain the panel recommended cut to maintain a higher performance expectation for students in the eighth grade. For a summary of the panel's main discussion points, see Appendix I: Vertical Articulation Panel Discussion .

DLM Recommended Cut Points and Impact Data

DLM staff accepted the recommendations made by the vertical articulation panel and recommended those cut scores for all subsequent reviews made by the TAC and DLM science states. That is, DLM staff recommended the acceptance of the panel-recommended (raw) cut point for the sixth grade Emerging/Approaching cut and the statistically adjusted cut points for all other cuts. DLM staff further recommended the acceptance of the panel-recommended cut points for end-of-instruction biology. Table 12 and

Table 13 below display the full set of the DLM-recommended cut points and associated impact data, respectively. The panel-recommended cut points were carried forward as the DLM staff recommended cut points. **Error! Reference source not found.**Figure 3 summarizes the percent of students in each performance level for each grade based on the DLM cut point recommendations.

Table 12. DLM-Recommended Cut Points for Science

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Required Linkage Levels
4	9	15	21	27
5	10	17	25	27
6	9	15	21	27
8	10	16	23	27
9-12	8	16	23	27
Biology	9	15	22	30

Table 13. Percentages of Students in Each Performance Level Based on DLM-Recommended Cut Points

Grade	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target/Adv (%)
4	59.4	24.0	12.6	4.0	16.6
5	58.5	21.9	15.4	4.2	19.5
6	45.4	30.6	19.2	4.8	24.0
8	52.6	25.5	18.7	3.2	21.9
9-12	54.1	29.6	13.0	3.3	16.3
Biology	32.3	20.0	22.3	25.5	47.8

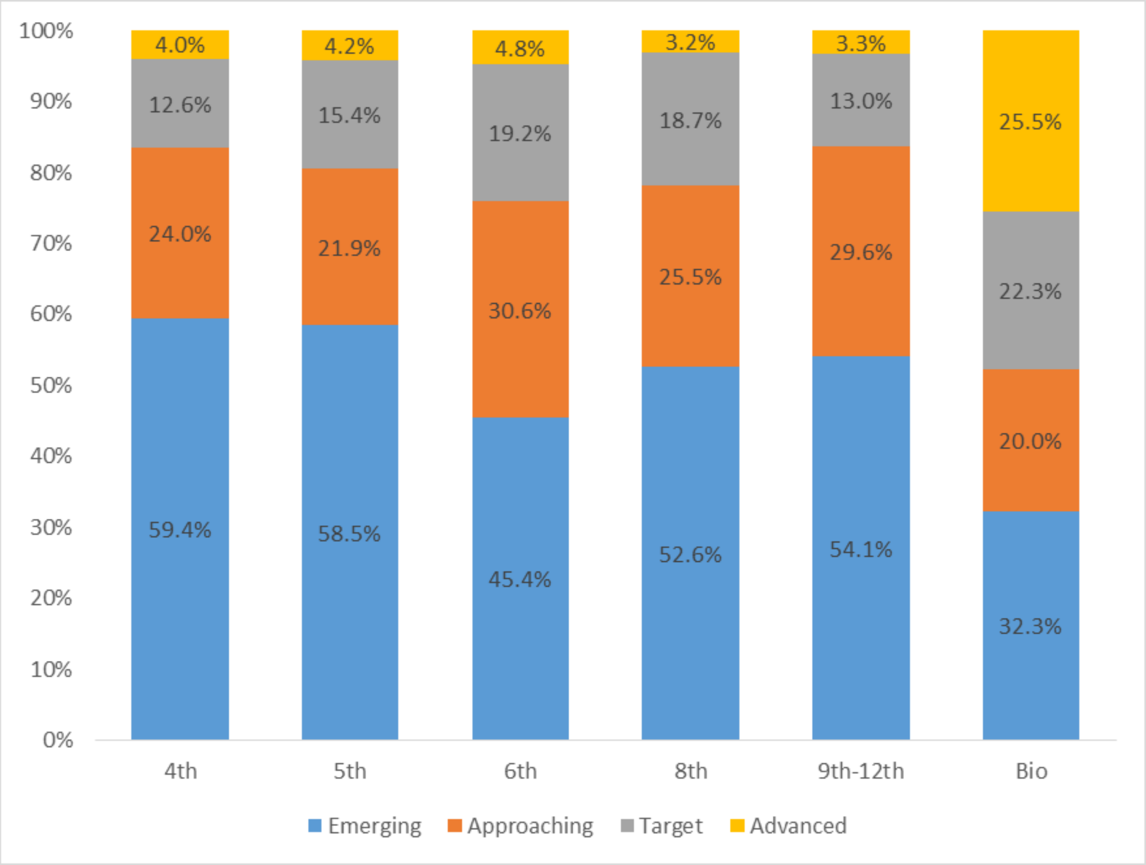


Figure 3. Impact Data Using DLM-Recommended Cut Points for Science

Evaluations

At the conclusion of the standard-setting meeting, panelists completed evaluations of the process. The questionnaire included panelist evaluation of the panel-recommended cut points, as well as their evaluation of the panel meeting process and overall feedback on their experience.

Independent Panelist Evaluations of Panel-Recommended Cut Points

As part of the evaluation process, panelists were asked to provide their final independent rating of the panel-recommended cut points. For each cut point, a scale of -7 to 7 was provided for the panelist to indicate how they would adjust the panel-recommended cut point. If the panelist agreed with the panel's recommendation, zero was circled, otherwise the panelist could indicate the value by which they recommended adjusting the value set by the panel. Table 14 summarizes panelist responses from their final independent rating of the cut points. Note that the percent included in the table is based on all three cut points. Panelists were asked whether they would choose to adjust the cut points three times: once for the Emerging/Approaching cut, once for the Approaching/Target cut, and once for the Target/Advanced cut.

Table 14. Panelist Comfort with Group Recommended Grade and EOI Cut Points

Grade	*N* Panelists	*N* Ratings*	*n* No Adjustment	Percent No Adjustment
4	4	12	10	83.3
5	4	12	12	91.7
6	4	12	11	100.0
8	4	12	12	100.0
9-12	8	24	24	100.0
Biology	8	24	24	100.0

Note: * = *n* Panelists × *n* Cut Points Evaluated

Across all panelists, panels, grades/courses, and cut points (*N*=96), 96.9% of panelists (*n* = 93) indicated that they would not choose to adjust the cut point. Only 3.1% of responses (*n* = 3) indicated that they would choose to adjust the group-recommended cut point. Complete panelist agreement with the recommended cut point was found in 16 out of 18 cuts (88.9%) across all grades and courses. There were three instances where a panelist indicated they would adjust the cut point if given the option: Grade 4 Emerging/Approaching, Grade 4 Approaching/Target, and Grade 6 Target/Advanced. In each instance, the indicated adjustment was -1 linkage level. Unanimous panelist comfort with all three recommended cut points was found for four out of six cut point panels (66.7%).

Panelist Evaluations of the Meeting

In addition to providing recommendations on the panel's cut points, panelists also evaluated the overall panel meeting process. The evaluation included self-evaluation of readiness to rate profiles, understanding of the tasks, and evaluation of outcomes. Panelists rated their responses to the 22 questions on a Likert scale, choosing either "Strongly Disagree" (SD), "Disagree" (D), "Agree" (A), or "Strongly Agree" (SA). For the last three questions, "Not applicable" was an additional option.

Table 15 shows that the majority of panelists agreed or strongly agreed that the meeting was well organized; they understood their tasks and felt confident to complete them, and they thought the cut points were defensible and valid. Furthermore, panelists believed that the meeting was a good experience in terms of professional development and for planning instruction with students with the most significant cognitive disabilities.

Table 15. Percentages of Science Panelist Responses to Evaluation Items

Question	SD	D	A	SA
1. The overall goals of the standard-setting panel meeting were clear.	0	0	5	27
2. The panel meeting was well organized.	0	0	4	28
3. The training and practice exercises provided the information I needed to complete my tasks.	0	0	5	27
4. It was clear what knowledge, skill, or ability a student would need to demonstrate to achieve a certain profile.	0	1	13	18
5. The profiles were representative examples of one or more of my students' knowledge, skills, and abilities.	0	1	7	24
6. Evaluating profiles was an effective way to set cut points for the performance levels.	0	0	5	27
7. I considered the *performance level descriptors* when I rated each profile.	0	0	7	25
8. I considered the *assessment items* when I rated each profile.	0	0	6	26
9. I considered the *other panelists' opinions* when I rated each profile.	0	1	6	25
10. I considered *my experience in the field* when I rated each profile.	0	0	6	26
11. I understood how to rate each profile.	0	0	9	23
12. I had enough time to complete the tasks.	0	0	3	29
13. I felt confident when rating the profiles.	0	0	9	23
14. The procedure for recommending cut points was free from bias.	0	0	5	27
15. Overall, I was satisfied with the ratings made by panelists in my group.	1	0	7	24
16. I am confident that the meeting produced valid cut point recommendations.	1	0	9	22
17. Overall, I believe my opinions were considered and valued by the group.	1	0	4	27
18. Overall, my group's discussions were open and honest.	1	0	3	28
19. Participating in the process increased my understanding of the DLM assessment.	1	0	3	28
20. Overall, I valued the panel meeting as a professional development experience.	1	0	2	29
21. This experience will help me plan and provide instruction for my students with significant cognitive disabilities.	1	0	3	27
22. This experience will help me use the DLM assessment more effectively.	1	0	4	26

Technical Advisory Panel (TAC) Evaluation of Panel Process

A member of the DLM TAC was on-site for the duration of the standard-setting event. The goal was to observe the process and provide feedback to the TAC and consortium state partners regarding any relevant observations of the event. Overall, the DLM TAC member believed that the standard-setting meeting was well planned and implemented, the staff were helpful to the panelists, and the panelists worked hard to set standards. The full TAC evaluated the evidence about the standard-setting process, including the TAC member's observations, panelist evaluations, and the relationship between panel and independent cut points. The TAC accepted the resolution about the adequacy, quality of judgments, and extent to which the process met professional standards. A copy of the memorandum and resolution is provided in Appendix H: TAC Resolution on DLM Standard Setting.

Final Results

The panel-recommended cut points, DLM-recommended cut points, and associated impact data for both sets of cut points were presented to the TAC and partner states for review. The TAC approved the DLM adjustment method and the process used by the standard-setting panelists and vertical articulation panel. Following the states' review process and discussion with the DLM team, the states voted to accept the DLM-recommended cut points as the final consortium cut points with no further adjustment.

Chapter 6: Future Steps

This technical report describes the steps in standard setting from developing policy-level PLDs through consortium adoption of cut points. Since the chosen standard-setting approach was student-based rather than item-based, grade-specific PLDs were not developed for use during the panel process. Instead, grade-specific PLDs will be developed from the work done by panelists as they evaluated profiles. Starting with raw notes about critical skills and understandings for each performance level and the associated rationales, DLM test development content teams will draft PLDs for each grade. These drafts will go through rounds of review and input from the partner states before they are finalized.

References

Bridgeman, B. (2013). *Human Ratings and Automated Essay Evaluation*. In M. D. Shermis & J. Burstein (Eds.), Handbook of Automated Essay Evaluation (221-232). New York, NY: Routledge.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement, 15*, 13-21.

Cizek, G. J., & Bunch, M. B. (2006). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 433-470). New York, NY: American Council on Education/Praeger.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). *Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results*. Setting performance standards: Foundations, methods, and innovations, 47-76.

Karvonen, M., Clark, A., & Nash, B. (2015). *2015 year-end model standard setting: English language arts and mathematics* (Technical Report No. 15-03). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Kingston, N. M. & Tiemann, G. C. (2012). *Setting performance standards on complex assessments: The body of work method*. In G. J. Cizek (Ed.). Setting performance standards: Concepts, methods, and perspectives (2nd ed.). New York, NY: Routledge.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Appendix A: DLM 2015-2016 Scoring Model Description for Science

Essential Elements (EEs) are academic content standards for students with the most significant cognitive disabilities. For each tested EE in science, assessments are available at one of three linkage levels that represent the relative progression toward the academic standard. For each part of the assessment, the student receives a testlet at a linkage level.

The DLM scoring model used for operational purposes in 2015-16 for science was constructed based on information obtained from students at each linkage level separately and then aggregated to produce student linkage level mastery estimates.

Students taking testlets at a linkage level within an EE were considered masters of that linkage level if one of two conditions were met:

1. The posterior probability of mastery determined from the diagnostic classification model estimated for the linkage level was greater than or equal to .80.
2. The proportion of items answered correctly within the linkage level was greater than or equal to .80.

Students were considered masters by meeting either condition in order to prevent consequences associated with false negatives. Linkage levels were treated hierarchically in that masters of higher linkage levels (based on the two criteria above) were automatically assumed to be masters of lower linkage levels. Students who did not demonstrate mastery at any linkage level were assumed to be masters of linkage levels at least two categories below the highest linkage level where they tested. Students who did not meet mastery criteria and whose highest level tested was either the Initial or Precursor levels were considered non-masters of all linkage levels.

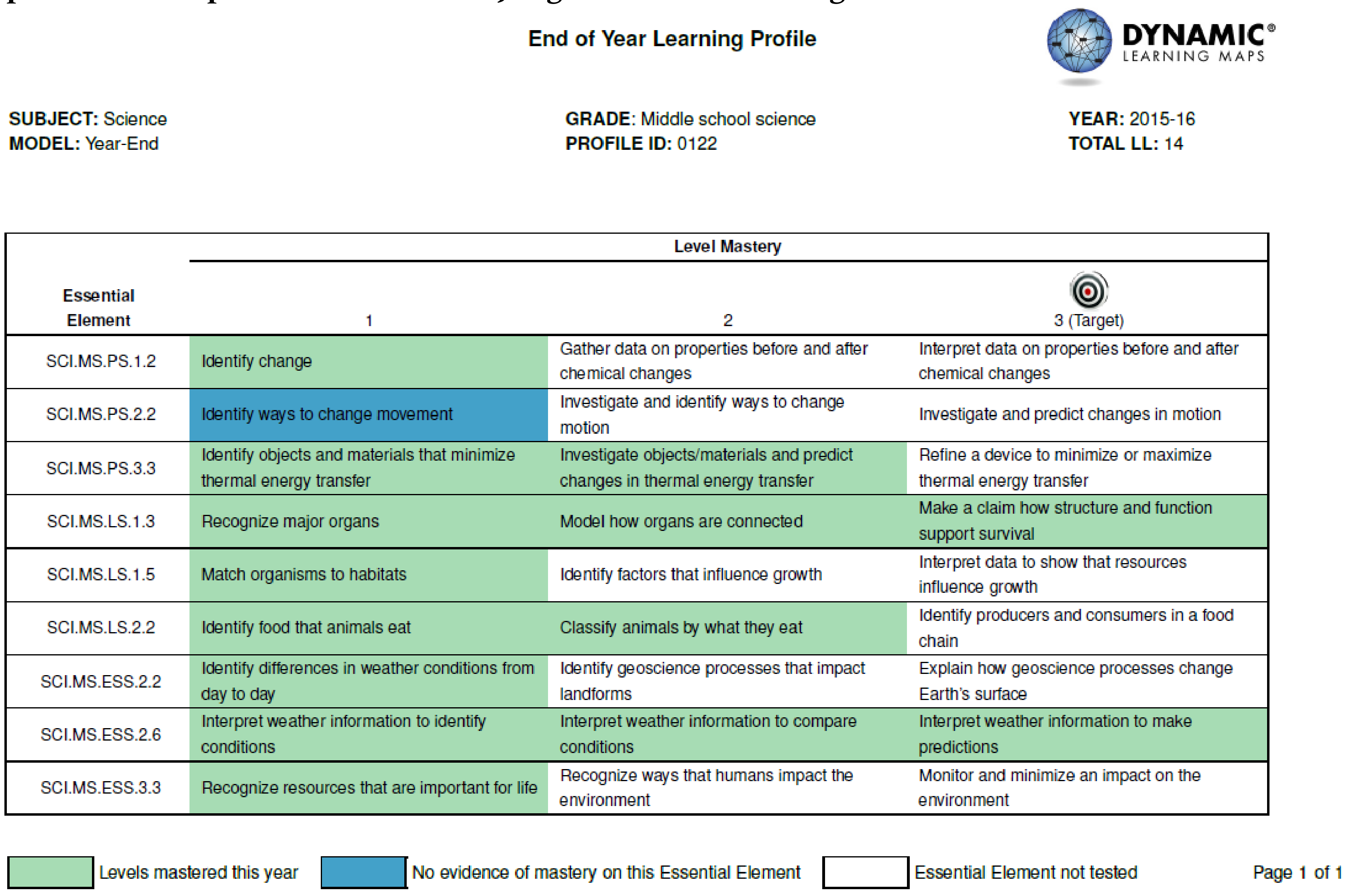
The diagnostic classification model used to classify students within each linkage level was the "Noisy Inputs, Deterministic Or gate" (NIDO) model (e.g., Rupp, Templin, & Henson, 2010; Templin, 2006). In this model, all items from each linkage level within each EE are treated as measuring one binary latent variable that represents mastery status for a student. All items within a linkage level are treated as exchangeable or fungible, a condition made necessary due to many items not being administered to large numbers of examinees. Fungibility (from the NIDO model) means that within a linkage level, all item parameters are constrained to be equal, providing the same item intercept and main effect parameters.

References:

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods ,and Applications.* New York: Guilford (Chapter 6, p. 135).

Templin, (2006). [Computer Program Manual]. CDM User's Guide. Lawrence, Kansas.

Appendix B: Sample Profile Based on Judgments about Linkage Levels: Middle School Science



End of Year Learning Profile

DYNAMIC®
LEARNING MAPS

SUBJECT: Science
MODEL: Year-End

GRADE: Middle school science
PROFILE ID: 0122

YEAR: 2015-16
TOTAL LL: 14

Essential Element	Level Mastery		
	1	2	◎ 3 (Target)
SCI.MS.PS.1.2	🟩 Identify change	Gather data on properties before and after chemical changes	Interpret data on properties before and after chemical changes
SCI.MS.PS.2.2	🟦 Identify ways to change movement	Investigate and identify ways to change motion	Investigate and predict changes in motion
SCI.MS.PS.3.3	🟩 Identify objects and materials that minimize thermal energy transfer	🟩 Investigate objects/materials and predict changes in thermal energy transfer	Refine a device to minimize or maximize thermal energy transfer
SCI.MS.LS.1.3	🟩 Recognize major organs	🟩 Model how organs are connected	🟩 Make a claim how structure and function support survival
SCI.MS.LS.1.5	🟩 Match organisms to habitats	Identify factors that influence growth	Interpret data to show that resources influence growth
SCI.MS.LS.2.2	🟩 Identify food that animals eat	🟩 Classify animals by what they eat	Identify producers and consumers in a food chain
SCI.MS.ESS.2.2	🟩 Identify differences in weather conditions from day to day	Identify geoscience processes that impact landforms	Explain how geoscience processes change Earth's surface
SCI.MS.ESS.2.6	🟩 Interpret weather information to identify conditions	🟩 Interpret weather information to compare conditions	🟩 Interpret weather information to make predictions
SCI.MS.ESS.3.3	🟩 Recognize resources that are important for life	Recognize ways that humans impact the environment	Monitor and minimize an impact on the environment

🟩 Levels mastered this year 🟦 No evidence of mastery on this Essential Element □ Essential Element not tested

Page 1 of 1

Appendix C: Standard Setting Panelist Recruitment Letter and Survey

Dear Colleagues,

[State] is a state partner in the Dynamic Learning Maps (DLM) Science Alternate Assessment Consortium. DLM science assessments are designed for students with the most significant cognitive disabilities and measure student mastery of content in science. The 2015-2016 academic year is the first year the DLM science assessment is operational. Student responses obtained during this first operational testing window will be used to determine what level of mastery is associated with certain performance levels. This process is referred to as standard setting.

As a partner state, we have the opportunity to recruit educators to serve on one of four panels that will help set standards:

- Elementary (grades 3-5)
- Middle (grades 6-8)
- High school (grades 9-12)
- High school biology (end-of-instruction 9-12)

We are writing to invite volunteers from [state or district] to serve on these four DLM standard-setting panels. We seek educators with a broad array of perspectives and backgrounds, although we especially seek individuals with content expertise in science and in education and assessment for students with significant cognitive disabilities. Other subject matter experts and individuals who work at establishments that employ individuals with significant cognitive disabilities are also encouraged to volunteer to serve on high school panels.

We ask that panelists commit to up to four hours of training in advance of the meeting and to attend and on-site standard-setting meeting in Kansas City, MO, June 15-17, 2016. Panelists must be present for the entire on-site meeting. Panelists who participate outside the scope of their usual job requirements will be paid a stipend of $600 to complete advance training and participate in the entire on-site meeting.

Volunteers are invited to complete a background survey online by following the link provided (https://kansasedu.qualtrics.com/SE/?SID=SV_bIZapjxIBg3xDql). The deadline to volunteer to participate in a standard-setting panel is Friday, April 8, 2016. DLM staff will notify volunteers who are selected to serve on panels.

We would appreciate your assistance with recruiting volunteers to serve as standard-setting panelists.

Questions about the standard-setting process should be directed to dlm@ku.edu.

Thank you for your assistance with the recruitment process!

Sincerely,

Intro DLM Standard Setting Panel Survey
Provided via Qualtrics

Please tell us about yourself and your interest in participating as a standard-setting panel member. Thank you!

Q1 First name

Q2 Last Name

Q3 E-mail Address

Q4 Preferred Phone Number

Q5 Full Mailing Address
Street Address 1
Street Address 2
City
State
Zip

Q6 What is your current role?

● Classroom Teacher
● Building Administrator
● District Staff
● State Education Agency Staff
● University Faculty/Staff
● Community Member
● Other ____________________

Q7 Please adjust the bars to indicate your years of p-12 educational experience in each of the following areas.
______ Science
______ Students with Significant Cognitive Disabilities
______ p-12 Education Overall

Q8 Which of the following types of students with disabilities have you taught/worked with in the past ten years? (Mark all that apply)

❏ Blind/Low Vision
❏ Deaf/Hard of Hearing
❏ Emotional Disability
❏ Mild Cognitive Disability
❏ Multiple Disabilities
❏ Orthopedic Impairment

❏ Other Health Impairment
❏ Severe Cognitive Disability
❏ Specific Learning Disability
❏ Speech Impairment
❏ Traumatic Brain Injury
❏ None of the Above

Q9 Which grade(s) did you teach in 2014-15?

❏ Grade 3
❏ Grade 4
❏ Grade 5
❏ Grade 6
❏ Grade 7
❏ Grade 8
❏ Grade 9
❏ Grade 10
❏ Grade 11
❏ Grade 12
❏ I did not teach in 2014-15

Answer If *Which grade(s) did you teach in 2014-15?* None Is Selected
Q9b Please indicate the grade band(s) at which you believe you have expertise to participate in standard setting.

❏ Grades 3-5
❏ Grades 6-8
❏ Grades 9-12

Q11 How many years of experience do you have teaching at these grade levels?
______ Years of Experience

Q12 Do you have previous experience with a standard setting process for another large-scale assessment besides DLM assessments?

● Yes
● No

Q13 How many years of experience do you have with Alternate Assessments based on Alternate Achievement Standards (AA-AAS)?

● None
● Less than 1 year
● 1-5 years
● 6-10 years
● 11+ years

Q14 Have you written items for DLM assessments?

● Yes
● No

Q15 Have you previously served as an external reviewer for DLM assessments?

● Yes
● No

Q16 Please list all licensures/certifications you hold.

Q17 Please check all of the following statements that apply to you.

❏ I have/had a leadership role in curriculum planning in my school or district.
❏ I have/had a leadership role in special education in my school or district.
❏ I have worked on my state's alternate assessment (e.g., scoring, range finding).
❏ I have written items for a statewide assessment.

Q18 What is your gender?

● Male
● Female

Q19 What is your ethnicity?

● Hispanic/Latino
● Non-Hispanic/Latino

Q20 What is your race? (Choose one or more)

❏ White
❏ Black/African-American
❏ Asian
❏ American Indian/Alaska Native
❏ Native Hawaiian/Other Pacific Islander

Q21 What state do you work in?

● AK
● CO
● IL
● IA
● KS
● MI
● MS
● MO
● NH
● NJ
● NC
● ND
● OK

- ● PA
- ● UT
- ● VT
- ● VA
- ● WI
- ● WV
- ● Other

Answer If *In which state do you work?* Other Is Selected
Q21b If "Other" was selected, please list the state in which you work.

Q22 Which best describes the population density in your school/workplace?

- ● Rural (population living outside settlements of 1,000 or less inhabitants)
- ● Suburban (an outlying residential area of a city of 2,000-49,000 or more inhabitants)
- ● Urban (city of 50,000 inhabitants or more)

Q23 Will you be able to commit to completing up to four hours of advance training prior to the on-site standard-setting meeting?

- ● Yes
- ● No

Q24 Will you be able to attend the entire on-site standard-setting meeting on June 15-18, 2015?

- ● Yes
- ● No

Thank you for completing the survey. DLM staff plan to notify volunteers who have been selected to serve on panels within 14 days after a recruitment phase ends.

Appendix D: Panel Training and Materials

Large file – separate attachment to be included in final document

Appendix E: Example Rating Forms for Range Finding and Pinpointing

DLM Standard Setting
Rating Form – Range Finding

Panelist ID: ___________ **Table ID**: _________ **Subject**: Science **Grade/Course**: 5th

			Round 1 Rating				Round 2 Final Rating			
	Profile ID	# LLs	EM	AP	T	ADV	EM	AP	T	ADV
1	0013	5								
2	0014	5								
3	0015	5								
4	0028	10								
5	0029	10								
6	0030	10								
7	0043	15								
8	0044	15								
9	0045	15								
10	0058	20								
11	0059	20								
12	0060	20								
13	0073	25								
14	0074	25								
15	0075	25								

DLM Standard Setting
Pinpointing Form: AP/T

Panelist ID: ___________ **Table ID**: _________ **Subject**: Science **Grade/Course**: _____________

			Round 1 Rating				Round 2 Final Rating			
	Profile ID	# LLs	EM	AP	T	ADV	EM	AP	T	ADV
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										

EM = Emerging AP = Approaching Target T = At Target ADV = Advanced

Appendix F: Panelist Meeting Evaluation Form

Table ID: ________ **Subject**: Science **Grade/Course: ____________** **Panelist ID: _________**

Dynamic Learning Maps Science Standard Setting Panelist Questionnaire

June 2016

I. Panel Meeting Evaluation

Please consider the statements below and place an "X" in a box to indicate the level of agreement or disagreement you have with each statement. A rating scale ranging from strongly disagree to strongly agree is provided. Please mark only one of the options for each statement.

	Strongly Disagree	Disagree	Agree	Strongly Agree
1. The overall goals of the standard-setting panel meeting were clear.				
2. The panel meeting was well organized.				
3. The training and practice exercises provided the information I needed to complete my tasks.				
4. It was clear what knowledge, skill, or ability a student would need to demonstrate to achieve a certain profile.				
5. The profiles were representative examples of one or more of my students' knowledge, skills, and abilities.				
6. Evaluating profiles was an effective way to set cut points for the performance levels.				
7. I considered the ***performance level descriptors*** when I rated each profile.				
8. I considered the ***assessment items*** when I rated each profile.				
9. I considered the ***other panelists' opinions*** when I rated each profile.				
10. I considered ***my experience in the field*** when I rated each profile.				
11. I understood how to rate each profile.				
12. I had enough time to complete the tasks.				
13. I felt confident when rating the profiles.				
14. The procedure for recommending cut points was free from bias.				

In the space below, please feel free to:

- Add comments regarding any of the responses to the questions above
- Make suggestions to improve future standard setting workshops
- Tell us what you liked and/or did not like about the workshop

II. Cut Point Evaluation

Indicate your final, independent recommendation for each of your panel's recommended cut points.

- If you agree with the panel recommendation, circle 0.
- If you disagree with the panel recommendation, circle a number above or below 0 to indicate the direction and distance away from your panel's recommendation where you believe the cut point should be set.

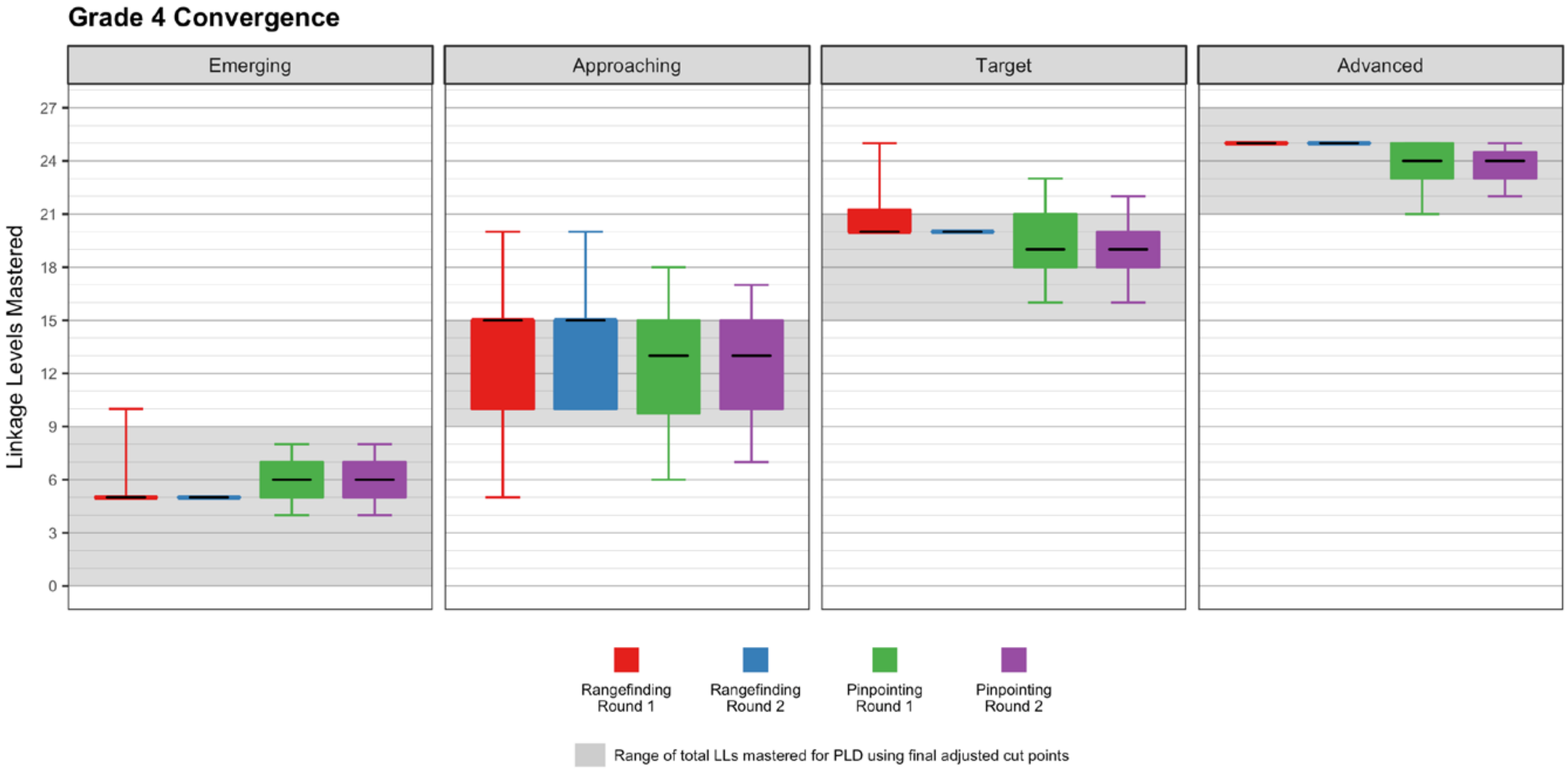
EM/APP	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7
APP/T	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7
T/ADV	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7

III. Overall Evaluation

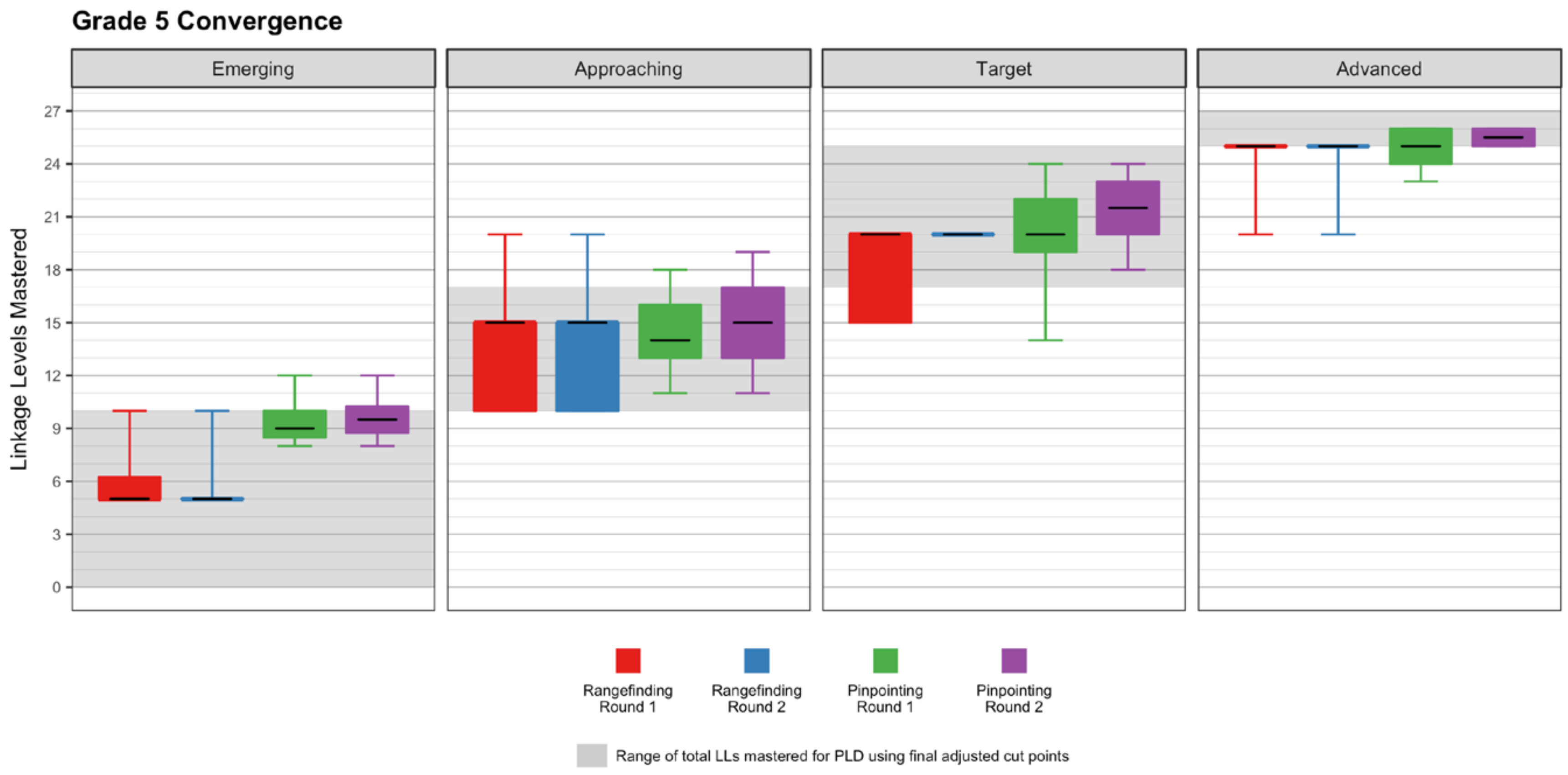
Please consider the statements below and place an "X" in a box to indicate the level of agreement or disagreement you have with each statement. Please mark only one option for each statement.

	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
1. Overall, I was satisfied with the ratings made by panelists in my group.					▨
2. I am confident that the meeting produced valid cut point recommendations.					▨
3. Overall, I believe my opinions were considered and valued by the group.					▨
4. Overall, my group's discussions were open and honest.					▨
5. Participating in the process increased my understanding of the DLM system.					▨
6. Overall, I valued the panel meeting as a professional development experience.					
7. This experience will help me plan and provide instruction for my students with significant cognitive disabilities.					
8. This experience will help me use the DLM system more effectively.					

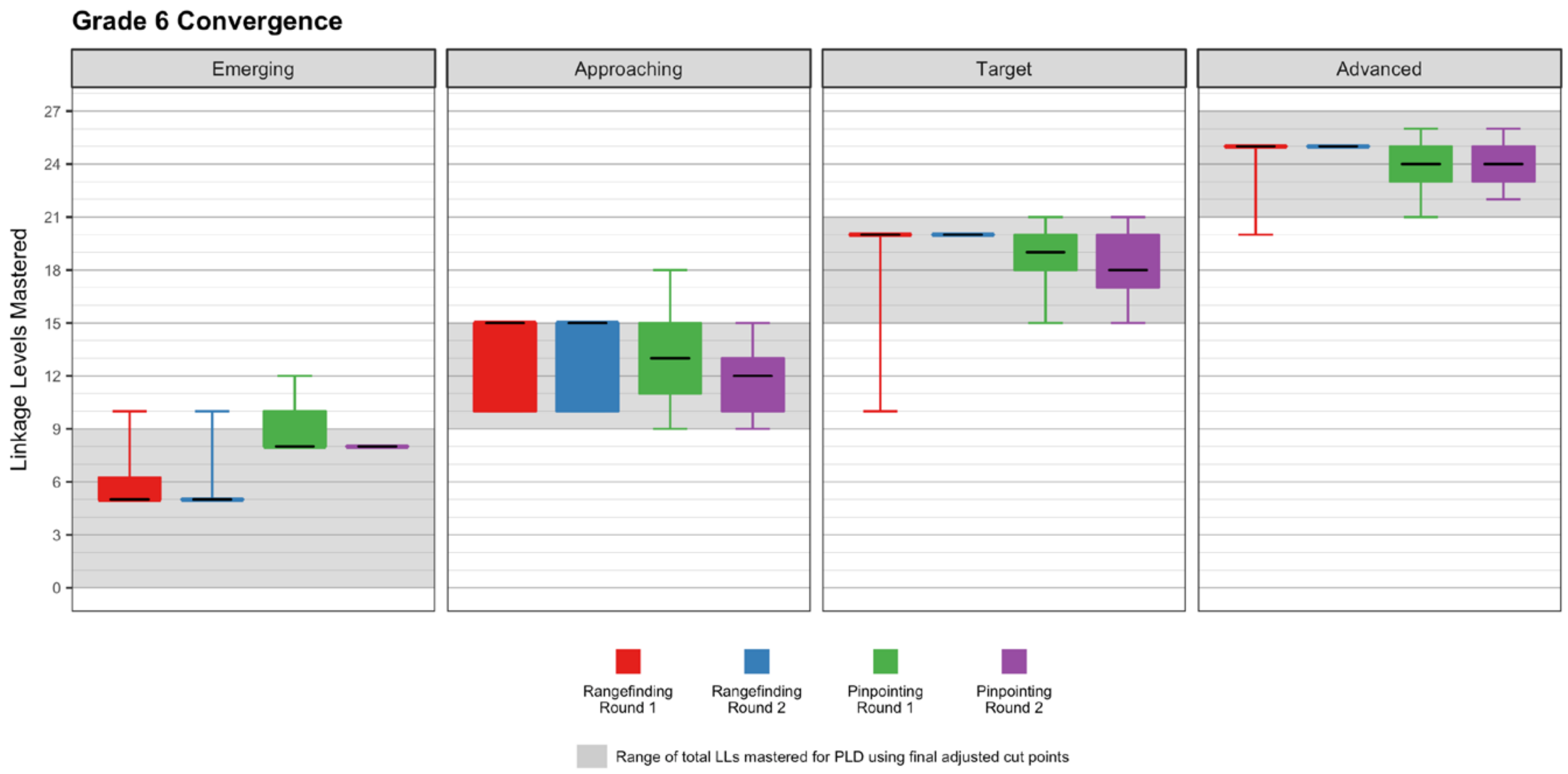
Appendix G: Convergence Plots for Range-Finding and Pinpointing Ratings



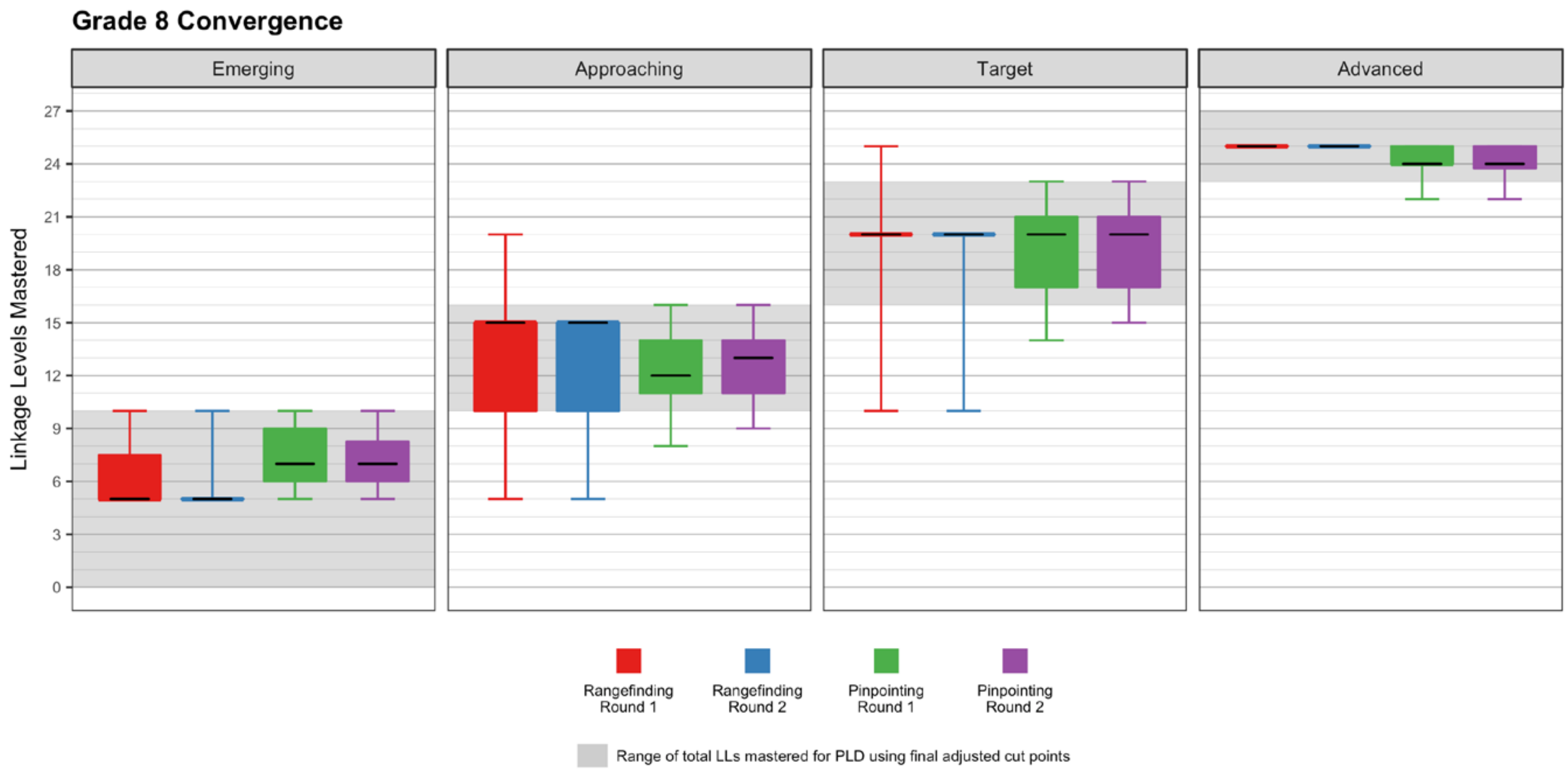
Note. The cut points represent the lowest value included in the higher performance level. For example, a cut point of 9 means that a LL mastery of 9 or greater is considered Approaching.



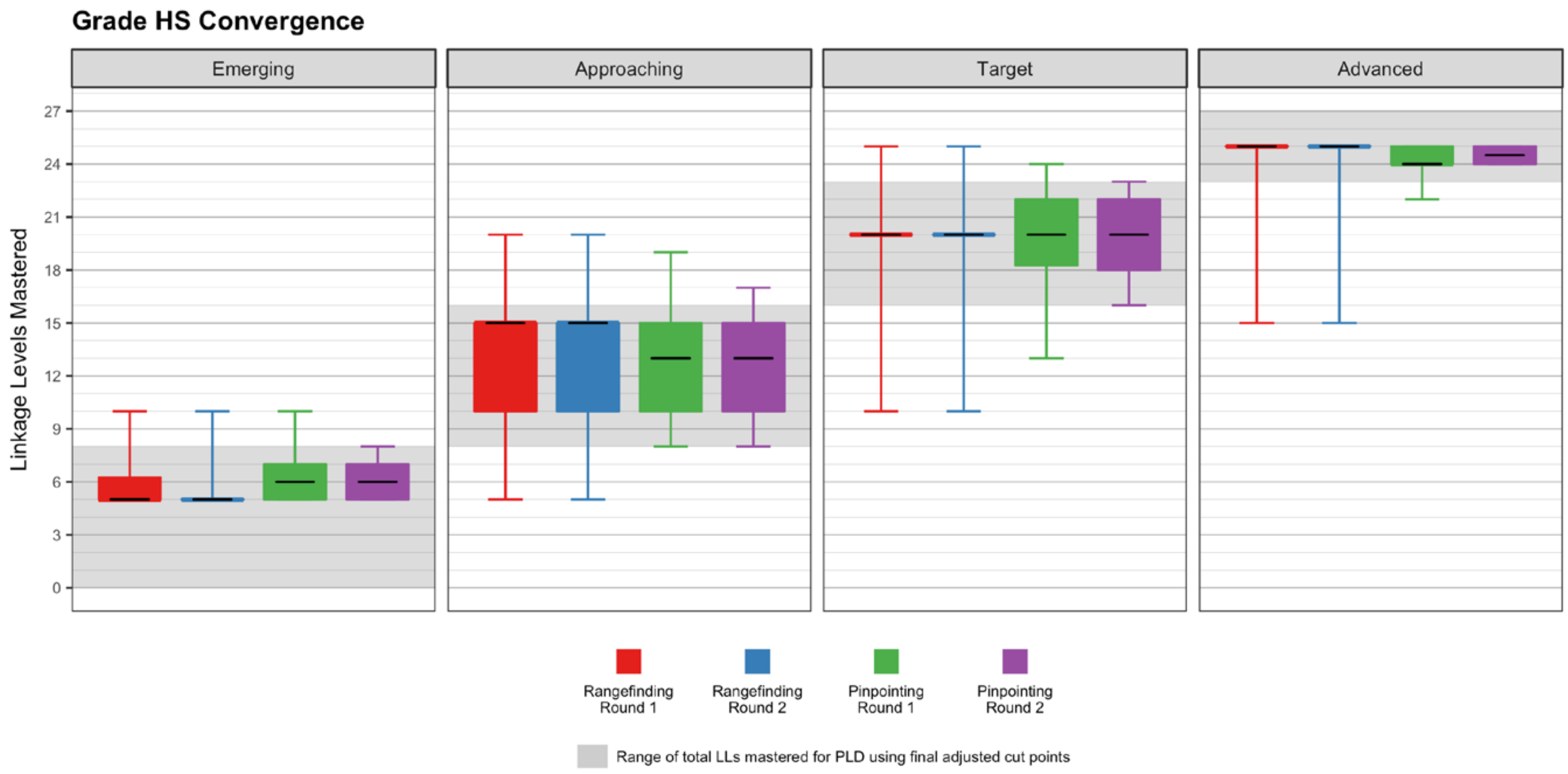
Note. The cut points represent the lowest value included in the higher performance level. For example, a cut point of 9 means that mastery of nine or more linkage levels is considered Approaching.



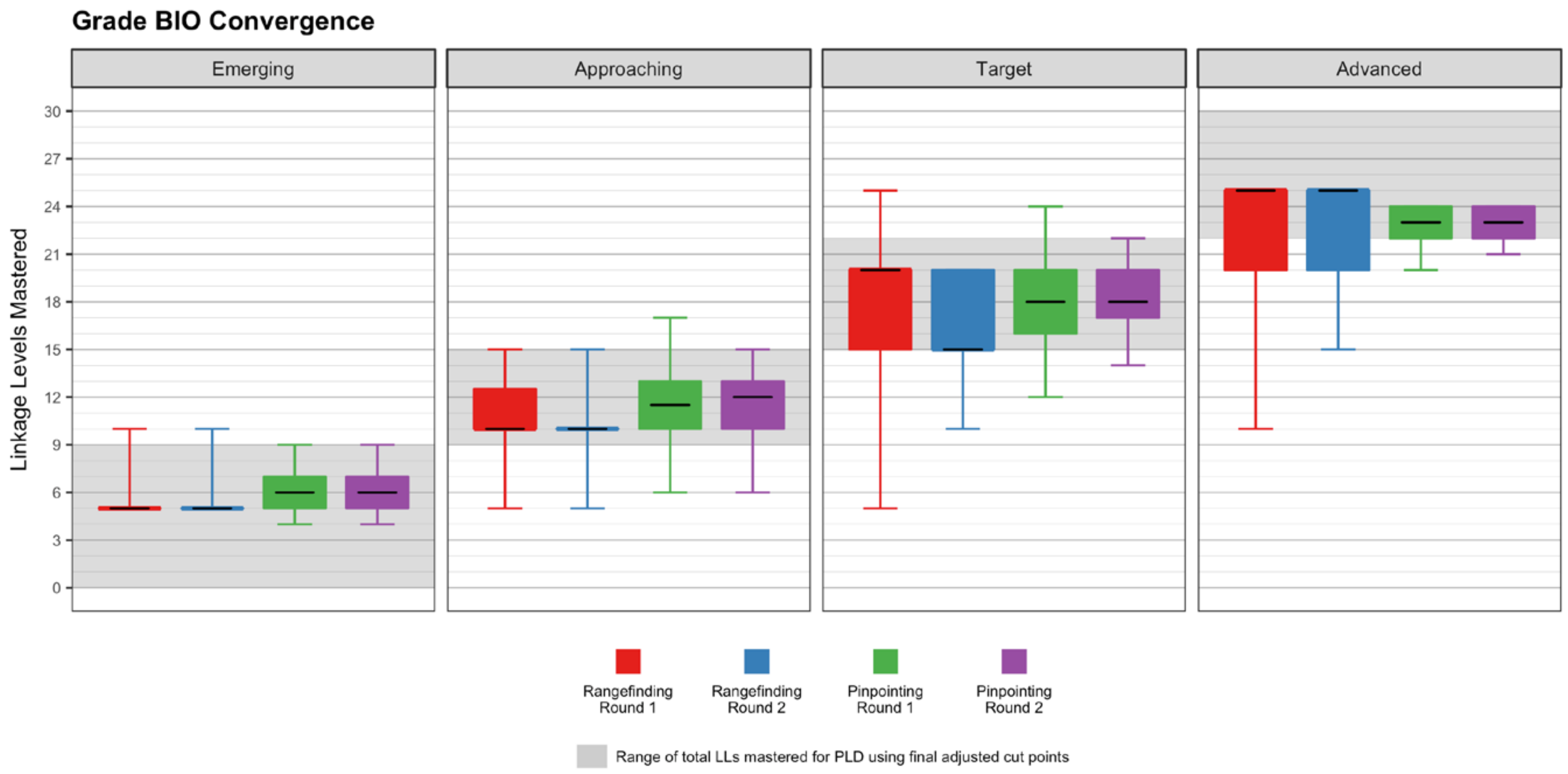
Note. The cut points represent the lowest value included in the higher performance level. For example, a cut point of 9 means that mastery of nine or more linkage levels is considered Approaching.



Note. The cut points represent the lowest value included in the higher performance level. For example, a cut point of 9 means that mastery of nine or more linkage levels is considered Approaching.



Note. The cut points represent the lowest value included in the higher performance level. For example, a cut point of 9 means that mastery of nine or more linkage levels is considered Approaching.



Note. The cut points represent the lowest value included in the higher performance level. For example, a cut point of 9 means that mastery of nine or more linkage levels is considered Approaching.

Appendix H: TAC Resolution on DLM Standard Setting

MEMORANDUM

To: DLM Staff and Participating States

From: Greg Camilli, member
DLM Technical Advisory Committee

Date: August 17, 2016

Subject: TAC Overview and Commentary on the DLM Science Standard Setting Process

As the representative of the DLM TAC, I was in attendance during the entire meeting conducted to set standards on the DLM science assessment which was conducted from June 15 through June 17, 2016. I provide observations below of the standard-setting process. These observations were shared with both the full TAC at the June 22 conference call as well as with the state members at their bi-monthly partner conference call.

The science assessment system follows the year-end model, which has a consistent blueprint that is covered in its entirety in the spring testing window. Assessments are available in grade spans (3-5, 6- 8, HS) and EOI Biology. Based on recommended TAC feedback and science states' input, cut points were set at fourth, fifth, sixth, and eighth grades as well as high school and high school biology. These are the specific grades in which DLM science states currently test for accountability purposes.

Overview of the Standard Setting Process for Science

1. The basic method of standard setting previously applied to ELA and mathematics was modified for science. The two main modifications were (1) the inclusion of impact data, and (2) a discussion with a cross-panel group of modifications for potential modifications to either raw or adjusted cut points.
2. Six panels of educators were convened by DLM staff representing fourth, fifth, sixth, and eighth grades as well as high school and high school biology.
3. Each panel had 4 to 8 members and set cut points for one of the 6 levels.
4. Considerable pre-meeting training (two-and-a-half hours) available by internet was required of all participants. Pre-conference training was delivered via video presentation. I considered the videos to be effective. I did not observe any panelists who seemed uncomfortable or unfamiliar with the procedures based on participation on group discussions. A key part of training was to orient panelists to EEs, linkage levels, and testlets corresponding to LLs. In particular, they were asked to form an understanding of what kinds of items and responses correspond to each EE/LL combination. Also incorporated was the general framework of policy-level PLDs to

anchor an understanding of performance. Supporting reference materials included notebook with glossary, blank profile forms, and tables listing and describing EEs.

5. A training folder was provided with 6 profiles for identifying performance level prior to actual range finding. Panelists were debriefed on training and allowed time for discussion of the process and for addressing questions.
6. The actual standard-setting event was carefully scripted. The facilitators were familiar with the procedures based on the previous standard settings in ELA and mathematics. Scripts were available for training, range finding, pinpointing, and recommended cut points and consideration of impact data. However, the impact data could only influence cut points based on cross-panel discussion, which occurred after individual groups had completed their work. Group facilitators collected and recorded all judgment data which was verified by one panelist. Facilitators were trained in the content of the standard setting as well as group dynamics.
7. I found that final recommendations were primarily driven by content rather than impact data.
8. The standard-setting meeting was carried out effectively and all groups finished early on the third day.

Day 1

1. The meeting began with introductions of the DLM staff present, the state observers, and the TAC observer. A signed test security and confidentiality statement was required. Meeting logistics were detailed, and required paperwork was distributed and collected.
2. Initial training provided on day 1 was designed to serve as a refresher of the online training each participant had gone through. Panelists had no questions regarding the process. A brief psychometric overview of the assessments was given. This type of presentation would benefit greatly from more intuitive explanations (e.g., what does the "probability of mastery represent") and more effective graphics. A technical overview in 30 minutes is a difficult task.
3. Panels were instructed what would not be provided—no raw scores and no scale scores. The standard setting was grounded in terms of *total linkage levels*. One panelist questioned whether number of linkages levels was being used in lieu of a total score. Training activities included sample diagrams and sample student profiles. The linkage levels (LLs) mastered (that is, assessed and mastered) were shown in shaded green. LLs on which students were assessed but did not show mastery (Mastery is through either 80% correct or through the DCM mastery probability > .80 threshold) were shaded blue. The learning profiles that did not have any shading for linkage levels indicated no evidence of mastery. Some profiles also showed evidence that students only partially completed the assessment (i.e., rows for some EEs were entirely unshaded).
4. The initial instructions for interpreting blanks were somewhat confusing. This confusion was resolved on day 2 to the satisfaction of the panelists. In essence, the directive to "ignore blank cells" was amended to "blank cells do not provided evidence of mastery even if other LLs suggest mastery is plausible." Panelists were instructed

that four performance levels would be established—Emerging (EM), Approaching Target (AT), Target (T), and Advanced (ADV).

5. The steps in the process of setting standards were reviewed—two rounds of range finding with 10 student profiles that ranged from 5 to 25 in steps of 5, followed by two rounds of pinpointing with 21 profiles for each round (3 profiles for each of 7 LLs). Each round of pinpointing considered adjacent levels for each of three cut points. I observed panelists referring to PLDs, linkage level descriptors, and actual items at each LL in an iterative process.
6. Panelists were instructed to use the answer to this question to set standards: "Using your best professional judgment and considering all students with significant cognitive disabilities, which performance level best described this profile?"
7. Panelists were told that cut scores would be set by the number of LLs mastered. The number of LLs in range finding went from 5 to 25.
8. Panelists' ratings for range finding were indicated by panelists raising their hands to self-report their ratings and a summary tally with verification for each performance level. Data were entered into a predefined spreadsheet that contained the student profile number and profile scores (that corresponded to paper profiles prepared for each panelist in a group). The spreadsheet was projected on a wall for ease of viewing for the panelists. In round 1 of range finding, the scores entered served to trigger an indication in a separate table as to whether the level of agreement or disagreement warranted further discussion. Panelists were instructed to focus on these, as well as any other ratings that they wanted to discuss. During this activity, facilitators pointed out areas of discrepancy with regard to panel classifications as well as the vertical articulation of EM, AT, and ADV cut points. Panelists were reminded they were not required to agree on their judgments.
9. Once the round 1 ratings had been discussed, the panelists were instructed to enter their round 2 ratings. This resulted in a calculation of a suggested cut point. Based on the results of round 2 of range finding, a new set of profiles was provided to each group. To determine cut points, a logistic regression procedure was programmed into the spreadsheet.
10. I did see some disagreement about cut points, but this disagreement was primarily content based, and led to further discussion of key skills. Panelists were asked to classify profiles into proficiency groups independently and without discussion. (This was generally the case for each round of ranging finding and pinpointing.) However, it was mentioned that a consensus was desirable based on group discussion and presentation of rationales.
11. Pinpointing also consisted of two rounds using a potential LL range of 2 to 27 (2 to 30 in Biology). The pinpointing results (the cut points suggested by the panel) differed from range-finding results primarily for the EM/AT cut point. This the pinpointing step appears to be a necessary component of the standard-setting procedure. All panels had completed range finding by the end of the first day.
12. The issue of "blank space interpretation" was covered the staff debriefing on June 15, 2016, and a plan was devised for addressing panelist confusion.

Day 2

1. At the start of day 2, DLM staff addressed the “blank space issue” with the full group. Following some discussion, it appeared that panelists were able to understand the original intent.
2. The primary activity of day 2 was to complete pinpointing, to review impact data, and to identify key skills for performance levels. Panels were provided cut point ranges by DLM psychometricians (to avoid suggesting a particular cut point), and runners then provided panels with pinpointing folders that included additional profiles tailored to the cut point ranges and pinpointing forms.
3. In some panels, a few sentences were written to describe each of the performance levels as a precursor to grade-level PLDs: specifically, the KSAs addressed at each performance level. Other panels prepared a list without summary prose. I was informed that the DLM staff would take bulleted skills and prose and develop these into statements. This activity occurred primarily at the end of pinpointing. Near the beginning of day 2, panels were informed that skill identification would be a key task for facilitators, and they should refer to their notes and other materials for this discussion.
4. At the end of day 2, the staff debrief covered timing logistics. Most panels had completed or nearly completed the skill identification for proficiency levels. It was decided to provide all panels a brief amount of time to complete and review this activity at the beginning of day 3.
5. A special procedure was devised for the potential result that lower grades had higher cuts points than higher grades. However, this inconsistency did not emerge at the end of range finding and pinpointing.

Day 3

1. After finishing identification of key skills at the beginning of day 3, two members of each panel were identified to form a cross-panel, vertical articulation group for the purpose of evaluating cut scores set by the group as a whole. While this group met, the remaining panelists were debriefed, including DLM staff expressing appreciation of their work. For the vertical articulation panel, panelists were shown final cut points, adjusted cut points, and impact data.
2. Panelists were asked to consider the raw cut points and cut points smoothed across grades. Then they were asked if they were in agreement with those cut scores, and if not, what their cut score recommendations would be.
3. Cross-panelists were asked to focus on the following questions:
    - Do the percentages of students in each category roughly match what you would expect, based on your knowledge of students with the most significant cognitive disabilities?
    - What might explain the distributions you see here?
    - Do you believe the recommended cut points are reasonable, from content and policy perspectives?

- If you believe changes are needed:
    - o Where are changes needed?
    - o What is your rationale for making those recommendations? Content? Policy?
    - o What would be the impact on content at those performance levels?

4. In general, panelists recognized that the difference between raw and adjusted cut points could reflect a reasonable amount of disagreement about skills that were essential to a performance level. This led one panelist to remark that consideration of cut points and impact could lead to refinement of performance level skills.
5. There was stronger recommendation for keeping the raw cut for EM/AP at sixth grade and more moderate interest in changing the eighth grade EM/AP cut. The rationales were based on the panelists' discussion of what originally drove their raw cut point recommendations.
6. At the completing of the cross-panel work, this panel was given the same debriefing mentioned above. This debriefing covered confidentiality requirements regarding what panelists were allowed to say about the process, meeting materials, cut points, and impact data. The procedure for submitting travel expenses was also explained.
7. Panelists were then asked to complete all questions of the standard-setting evaluation questionnaire. This assessment included items regarding their comfort level and understanding of the procedure, individual evaluation of cut points, and overall impressions. These results will be compiled and included in the full memo to the governance board.
8. This work was completed prior to noon on day 3. In the remaining time, panelists were asked to contribute to content issues regarding the assessment and instructional materials.

Commentary

1. The actual standard-setting event was carefully scripted. The training of the six facilitators who led the work at each panel's table included a full-scale tryout of the standard-setting process (i.e., actually setting standards based on sets of the materials that would be used at the event). This procedure provided detailed understanding of the standard-setting process and permitted all panels to receive the same instructions at each step in the process for each grade/course for which standards were set.
2. There were daily debriefs with the facilitators, which permitted any needed mid-course corrections to be made to the process or instructions. This served to keep the standard setting on schedule.
3. In the student profiles, cells were blank for LLs when the student did not test on the EE. Some panelists started to evaluate those empty cells compared to adjacent mastered LLs and believed the student should have mastered the blank cells. Panelists were retrained to focus on the cells that were shaded as part of their evaluation.
4. Changes that were recommended during the cross-panel discussion were based on the assessment content and the standards and were less influenced by the impact data that had been presented. The cross-panel discussion provided key insights to the final

cut points. Panelists agreed that many cut points could have gone in either direction by a point or two, and all panelists indicated the final cut points were acceptable.

5. The standard-setting meeting was carried out effectively, the staff were helpful to the panelists, and the panelists worked diligently to set standards. The panelists were very supportive of the processes they used to set standards.

Resolution

At the June 22 Technical Advisory Committee (TAC) meeting, the TAC evaluated the methodology and process that was used to determine cut points rather than the cut point values themselves. Using this criteria, the TAC found the process to be consistent with the proposed methodology. Additionally, the TAC stated they could find nothing that should prevent the states from accepting the cut scores. The TAC further recommended that when presenting this information to states at the governance meeting, additional information should be included in the report, including a more explicit explanation of which students were included in the impact data and how off-grade testers (students who test in grades that did not receive grade-level cut points, i.e., third and seventh grades) were handled, as well as recruitment procedures and demographics of the panelists.

Appendix I: Vertical Articulation Panel Discussion Summary

The vertical articulation panel, comprised of 10 members (two from each panel) met to review cut points and impact data. They evaluated information based on panel-recommended (raw) cut points and on statistically adjusted cut points. Panelists evaluated whether the cut points were logical across grades and whether they were appropriate based on the content and their states' policy perspectives. The panel also discussed whether they would recommend any changes to the raw or adjusted cuts and their rationales for those changes.

Before being shown any results, the panel was asked what patterns they would expect to see in the cut points and impact data across grades. There was general consensus that the panel expected a general increase in cut points from lower to higher grade levels. They expected the impact data to show higher achievement in the lower grades and lower achievement in the upper grades. Much of the discussion about their rationale for this expectation focused on students' opportunity to learn. Panelists indicated that students in upper grades had less exposure to the science curriculum than those in lower grades. These representatives from the grade-level panels also noted that their panel-recommended cuts reflected standards that were higher than what was being taught in classrooms. While they expected to see low performance based on 2016 impact data, they believed that over time and with more effective instruction, more students would reach the At Target level.

When presented with the panel-recommended and adjusted cut points, vertical panel representatives indicated that in general the patterns of cut points were as expected, perhaps even more consistent across grades than they expected. However, in reviewing the statistically adjusted cuts the panelists noted the lack of progression from sixth to eighth grades and explained that with two additional years of instruction, the eighth grade cut should be higher than the sixth grade cut point. It was determined that moving the sixth grade cut point down a point rather than increasing the eighth grade cut was more reasonable given the difficulty of the content.

Panelist views of the impact data were that they were reasonable for the first year of administration of the assessment. They did not expect to see large percentages of students at the Advanced level. They again commented on the need to "set the bar high" for students, and that while there were currently large proportions of students at the Emerging level, they expected the performance level distribution to shift upward over time.

The panel's final recommendation was to adopt the statistically adjusted cut points, with one exception: retain the panel-recommended sixth grade cut point between Emerging and Approaching (9) rather than the statistically adjusted cut point (10).