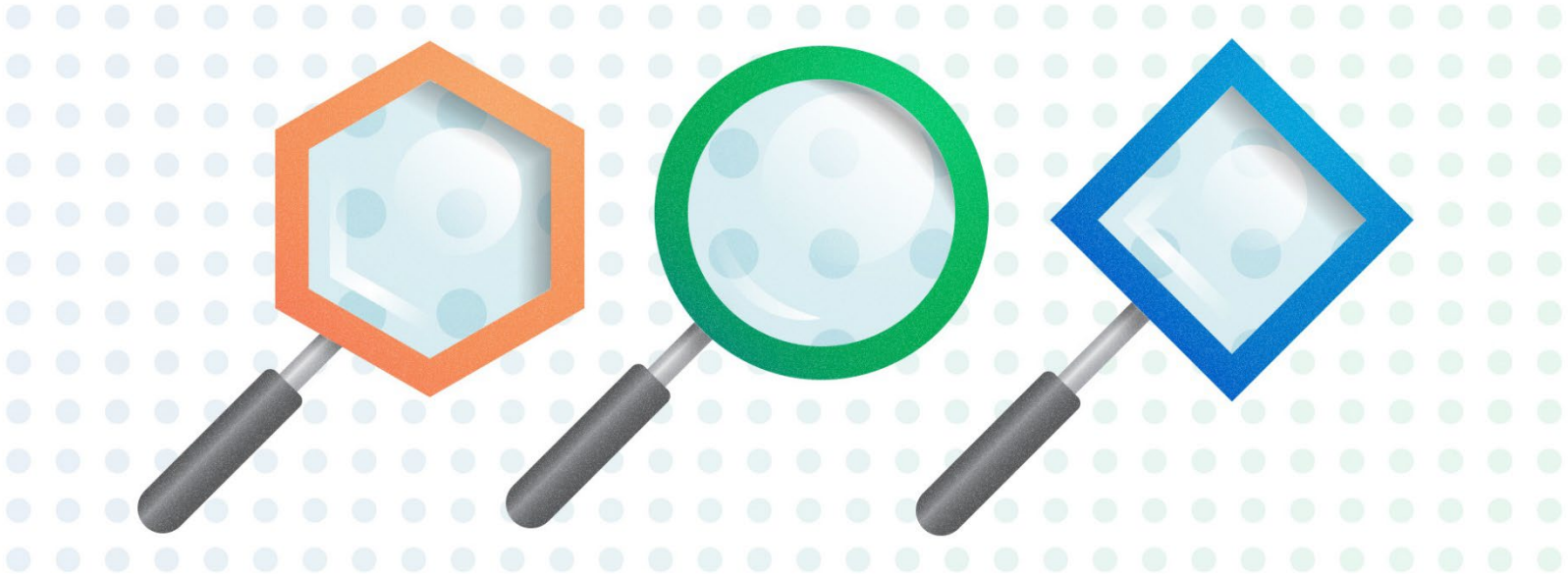

MONTANA JULY 2025 MAST STANDARD SETTING: OBSERVATION REPORT

August 2025

Will Lorié, Ph.D.
Senior Associate
Center for Assessment

Center for Assessment
National Center for the Improvement of Educational Assessment
Dover, New Hampshire

Table of Contents

Montana July 2025 MAST Standard Setting – Background

New Meridian, under contract with the Montana (MT) Office of Public Instruction (OPI), conducted a standard-setting workshop for Montana Aligned to Standards Through-Year assessments (MAST) for English language arts (ELA) and mathematics in grades 3 through 8. The workshop was held Monday, July 28, to Thursday, July 31, 2025, at Bozeman High School in Bozeman, MT.

The purpose of the workshop was to establish cut scores for the state's MAST assessment in grades 3 through 8 for ELA and mathematics. The workshop included vertical articulation activities on the afternoon of the final day.

OPI reviewed and approved a standard-setting plan (New Meridian, 2025). The plan outlines the standard-setting method for the workshop, including materials, activities, and the roles of various participants. New Meridian planned for and implemented a modified Bookmark standard setting procedure (Lewis, Mitzel, & Green, 1996; Lewis et al., 2012; Mitzel et al., 2001).

Panels of educators, organized by grade band and content area, engaged in a structured sequence of activities that included experiencing test content, reviewing performance level descriptors (PLDs), and developing descriptions of threshold students. There were three groups each in ELA and mathematics: A group for grades 3 and 4, another for grades 5 and 6, and a third for grades 7 and 8. Participants rendered judgments using a modified Bookmark procedure across two grade-level iterations (starting at the lower grade). The workshop concluded with vertical articulation discussions to ensure coherence of achievement level expectations across grades.

OPI contracted with the National Center for the Improvement of Educational Assessment (Center for Assessment) to observe these standard-setting meetings. Will Lorié (the author of this report) was the Center for Assessment observer.

Participants

Several people attended the standard-setting workshop, including staff from the OPI assessment office, led by OPI's Director of Assessment, Cedar Rose. The OPI Superintendent, Susan Hendalen, provided opening remarks on Day 1. Program management, content development, and psychometric staff from New Meridian, led by Tim Walker, Ph.D. (Senior Psychometrician), ran the standard-setting workshops. Caroline Lang, Ph.D. (Director, Program Management) and Rania Rotou, Ph.D. (Director, Psychometrics and Research), provided oversight and support for the New Meridian team.

Although planners expected eight panelists per group (for a total of 48 panelists), there were two cancellations, resulting in two rooms with seven participants. OPI selected the participants to represent educators from across the state.

Other participants included facilitators comprised of New Meridian content developers and psychometricians. New Meridian data analysts provided technical support and computed feedback and impact data. New Meridian also provided content-area experts who answered questions about the test.

Observation Notes

Day 1: Monday, July 28, 2025

General Session

The standard-setting workshop began with a general session that featured welcoming remarks from Superintendent Hendalen, who emphasized the importance of the standard-setting work for the state's accountability and reporting systems. She highlighted the innovative nature of the MAST program and praised her team's efforts. Next, Cedar Rose introduced key staff from OPI.

Caroline Lang introduced the New Meridian staff and outlined logistical procedures for the week. Tim Walker followed with a presentation that focused on the workshop's goals. He addressed security and reminded participants that, although the workshop would result in recommendations rather than final decisions, their work would be influential in shaping the MAST assessment.

A slide displayed the workshop goals, the first of which was to "complete a process that allows … SMEs to make judgements about the knowledge, skills, and abilities (KSAs[1]) Montana students would demonstrate in order to meet each of the four achievement levels." The slide indicated that participants would propose MAST cut scores that categorize students into the four achievement levels of novice, partially proficient, proficient, and advanced. It stated as a second goal that participants would "review and discuss how the achievement levels articulate across grade levels and suggest possible adjustments to ensure good articulation."

[1] Throughout the workshop, facilitators referred to "knowledge, skills, and abilities," "knowledge, skills, and practices," and "knowledge, skills, abilities, and practices." These were used interchangeably.

Tim presented a visual of New Meridian's test development process and introduced the technology tools to be used throughout the workshop, including the Smartsheet Dashboard and Waypoint platform. Since not all participants were familiar with these tools, New Meridian mentioned that they would allocate time in their breakout rooms to help everyone set up with Smartsheet and Waypoint.

Breakout Rooms

At around 10:00 AM, the panelists moved to their assigned grade-span rooms for content-specific sessions. The initial activities ensured all participants were connected to Smartsheet and Waypoint. New Meridian staff provided instructions for accessing testlets and other materials, as well as entering information into New Meridian's system, offering on-demand technical support through a hallway "genius bar" model. About 3-4 participants received assistance at any given time.

In the ELA 5/6 room, as in other rooms, the facilitator reviewed the results of a readiness survey to confirm participants' understanding of the purpose of standard setting and their role in the process. Participants then completed their testlets (experiencing the test from a student's perspective) and moved into discussions of achievement level descriptors (ALDs).

At 11:30 AM, in various rooms, facilitators introduced an activity focused on identifying the KSAs that define performance at each of the four achievement levels. Participants independently reviewed detailed ALDs and reflected on the students they had taught. They recorded one KSA per sticky note, labeled with the appropriate standard, and then shared and discussed these at their tables.

Participants had questions about how to perform the KSA task. In some rooms (e.g., Math 5/6), facilitators encouraged participants to proceed without a modeled example. In others (e.g., ELA 5/6), facilitators modeled the KSA task. Ultimately, all rooms followed a similar process, identifying KSAs on sticky notes and placing them on an achievement level backdrop (which could be the whiteboard, tables, or sheets of paper). KSAs mentioned by participants included statements such as "accurately identify point of view [and identify the detail] and its impact on the story" (ELA 5/6), needing a "fully labeled number line" for a particular type of fraction-related task (Math 3/4), and "I think that a novice might just be able to identify an illustration" (ELA 3/4).

Across rooms, facilitators emphasized using "would" rather than "should" when formulating KSAs, to avoid setting standards too high (the danger of "should," as explained by facilitators) and instead focus on likely student performance.

After lunch, participants continued the KSA activity. In some rooms, clear distinctions across achievement levels emerged, such as "identify" at novice, "analyze" at proficient, and "evaluate" at advanced in the ELA 7/8 room. Facilitators encouraged aligning KSAs with ALD language. Discussions became more specific and content-focused, with examples connected to actual testlet items.

By mid-afternoon (3:00 PM), facilitators moved on to the next activity: defining threshold students. Participants were asked to focus on the "gray areas" between levels (for example, students on the borderline between partially proficient and proficient) and to describe their characteristics. Slides directed participants to reflect on what a threshold student can probably do and what they still find challenging.

In ELA 5/6, the facilitator explicitly linked the threshold activity to the earlier KSA work, and participants generated content-based descriptors, such as "interpreting and integrating visuals into understanding of the text," as an indicator of movement from proficient to advanced. In Math 7/8, I observed participants articulating nuanced differences among borderline students, such as distinguishing between being able to "*identify* rational numbers" from "*explain* their properties."

Across content areas, engagement was high, and most groups concluded the day with detailed profiles of threshold students for at least one transition point. In some rooms (e.g., Math 3/4), participants began reviewing KSAs placed at the border between levels to determine whether they should be included in the threshold profiles.

Facilitators responded effectively to participant questions, which included questions about test scoring and item difficulty.

Day 2: Tuesday, July 29, 2025

Day 2 of the MAST standard-setting workshop started with panelists returning to their grade-band content groups. The initial activities focused on reinforcing assessment security protocols. In the ELA 5/6 room, for example, the facilitator discussed the use of personal devices. She emphasized the risks of viewing secure materials on personal computers and asked participants to make sure they do not capture secure content through screenshots.

Facilitators clarified the structure and interpretation of the Ordered Item Booklets (OIBs), explaining that the difficulties were calculated based on responses from all students in the state. They described the OIB as consisting largely of novice items at the start, followed by partially proficient, proficient, and finally advanced items toward the end. A slide stated, "items in between

are where the cut scores transition from one classification to the next." The reason for the repeated appearance of multipoint items was also explained. In ELA 5/6, participants asked if the OIB was a test, and the facilitator clarified that it is a collection of items from across the testlets, designed to reflect the range of difficulty across the testlets.

Panelists received printed item maps and were asked to examine each item in order, discuss the KSAs required to answer them correctly, and to annotate their item maps accordingly. The specific instructions, posted on a slide, were:

- At your table, discuss the OIB items, starting at the beginning of the booklet.
    - Consider which KSAs each item likely measures and why it is more difficult than the items preceding it.
    - Enter comments in the Item Map (or in your OIB) about the KSAs required to answer the items correctly for future reference.

Panelist discussions took place at tables and were explicitly tied to the KSAs identified during Day 1 activities.

The morning OIB discussions were centered on alignment to standards and student expectations. Panelists shared their reasoning about KSAs and debated whether an item should be considered novice, partially proficient, proficient, or advanced. The facilitator encouraged consistent reference to the ALDs, particularly the threshold descriptors developed during the previous day. A discussion at one table

By mid-morning, panelists received training on the Bookmark standard-setting method. Facilitators walked through the process of placing a bookmark at the transition point where a hypothetical threshold student would have a two-thirds probability of answering correctly. The rationale for this method, including its relation to the ALDs and OIB item sequence, was explained. The instructions for bookmark placement were shown on a slide:

- Working independently, consider a hypothetical student who is at the Novice level.
- Would the Novice Student likely answer the item correctly (at least 2/3 probability, i.e., have a 67% chance of answering correctly)?
    - If you answer "Yes", go to the next page.
    - Repeat the process until you switch your answer to "No." The OIB page number you are on is your bookmark location for the Partially Proficient level. It should align with your definition of the Partially Proficient threshold student.

There were similar slides for partially proficient and for proficient.

Participants completed a brief quiz to verify their understanding, with everyone required to demonstrate a solid grasp of the method before proceeding. In ELA 5/6, the facilitator reviewed one quiz item that some participants missed, using the opportunity to reinforce key concepts.

After lunch, the panelists started Round 1 of the bookmark placement activity. Working independently, they recorded bookmark locations for each performance level threshold (e.g., from novice to partially proficient, and from partially proficient to proficient). Participants had to provide a written rationale for each placement, referencing threshold student profiles and KSAs as supporting evidence.

Following the first round of placements, facilitators presented summaries of bookmark placements and also impact data to show the consequences of the cut scores implied by the initial bookmarks. In Math 5/6, participants asked about the statistical properties of the items and how item location was derived. Facilitators explained that item difficulty estimates incorporated both item p-values and discrimination, and that these were used to position the items in the OIB. The conversation in that room did not reference the KSAs as much as in the others. Participants did not engage in dialogue with each other prior to Round 2, as confirmed by the minimal change in

participants' judgments between the two rounds. New Meridian addressed the issue by providing for a third round in this group.

In ELA 5/6, the facilitator highlighted the variation in panelists' judgments and led a discussion about participants' bookmark placements, asking for their reasons. "I see her point," one participant said, in response to another's rationale. "I also felt that the vocabulary in that passage was very difficult," she added. When asked if the impact data should match MTSS percentages, the facilitator said, not necessarily. When a participant suggested that an item's format (selected response) played a big role in considering many of the items novice, the facilitator said that "that has to triangulate with the ALDs, the threshold description, and the test." After reviewing an item together, participants agreed that a question might seem simple, but if the related passage is highly complex, it could raise the item's difficulty.

Participants identified items with KSAs that seemed lower than expected based on their position in the OIB. For example, in ELA 7/8, a participant asked what to do when "you think they can't do [item] 8, but they can do 9, 10, and 11." The facilitator said that one could reevaluate the items' KSAs or their alignment with ALDs or threshold descriptors. He also acknowledged that item placement might sometimes reflect statistical anomalies (due to low discrimination) and encouraged panelists to consider multiple sources when placing their bookmark.

By the end of Day 2, most panels had finished their first round of bookmark placements and were either preparing for or had already completed Round 2. The discussions I heard were focused on content, with participants showing greater comfort with the standard-setting methodology. Facilitators emphasized that future rounds would allow for refining their work by

incorporating both feedback and impact data. Participants were also asked to complete end-of-day reflections.

In a debrief, New Meridian recommended that, since not all groups might have had sufficiently rich conversations between Rounds 1 and 2 (for groups that reached Round 2), a third round was necessary in some cases. This was put into place the next day. Additionally, New Meridian believed that in the ELA groups, participants had given too much emphasis to the writing items, which are challenging but account for a small part of the test score. Therefore, the team decided to provide information about writing for the ELA groups on Day 3, and all ELA groups would complete a third round of judgments.

Day 3: Wednesday, July 30, 2025

ELA General Session

Day 3 began with a short general session for ELA participants. Tim Walker of New Meridian praised the consistency of results across groups, calling the alignment "remarkable" and "a good thing." He then acknowledged an important issue: "We, as New Meridian, did not do the best we could in communicating the role of writing in the score." He explained that although writing items are difficult and have received a lot of attention in groups, they make up only about eight percent of the ELA score. The rest comes from reading. Therefore, there will be a third round of bookmark judgments in ELA. Tim encouraged panelists to focus their judgments mostly on reading, which makes up the majority of the score.

Breakout Rooms

I observed Math 5/6 on the morning of Day 3. The facilitator guided the group through a structured activity to gather each participant's reasons for their Round 2 bookmark placements, their flexibility in moving the bookmarks, and their reasons for where they might place them after such moves. They did this individually, then the facilitator had them share their reasons in a "round robin" style. When sharing their rationales, there was a content-focused back-and-forth among participants, consistent with best practices for these discussions.

The Math 7/8 facilitator explained that in the end-of-day reflections from the previous day, participants said they felt comfortable with their cut points, so they did not move to a third round. The group had moved on to set standards for their second grade (grade 8).

Later that morning, in a side meeting with OPI and New Meridian, I reviewed the impact data for ELA since all rooms had completed their Round 3 judgments for first grade. When I asked about the effect of the morning session on writing, New Meridian mentioned that it influenced the advanced cut score, making the impact data more reasonable with not so few students in the advanced category.

By late morning, all groups were working on their second-grade tasks. Overall, the activities for second grade mirrored those of first grade, but they moved faster because participants were already familiar with the process. Facilitators and participants made cross-grade comparisons during their discussions, using the work from first grade to help develop their KSAs for second grade. For example, in ELA 3/4, the facilitator presented detailed ALDs showing differences between those for standards RL 3.1 and RL 3.2 and their grade 4 counterparts. He highlighted that the language was similar but indicated an increased expectation for students: from "ask and

answer" to "refer to explicit details," especially for those who are partially proficient when transitioning from RL 3.1 to RL 4.1. The group then used what they had done for grade 3 to build their KSAs for grade 4.

Similarly, in ELA 5/6, when participants discussed the characteristics of students transitioning between levels, they built on the language used in the earlier grade. This was a whole group activity, with the facilitator writing the emerging threshold student descriptions on the whiteboard.

By mid-afternoon, rooms continued to create KSAs, threshold descriptions, or review the OIBs for their second assigned grades.

Similar to first grade, some groups had threshold student descriptions explicitly written on the whiteboard, some had the descriptions defined by a set of KSAs physically placed at the edges of the levels on a common display, and in some groups, the descriptions were at tables and written out by participants individually.

During a debrief session at the end of Day 3, facilitators and leadership staff discussed how to handle questions about language in test items, concluding that such questions should be linked back to the content standards.

Day 4: Thursday, July 31, 2025

On the final day of the workshop, participants finalized cut score recommendations and engaged in vertical articulation across grades. The vertical articulation took place in two rooms, one for ELA and another for math.

Breakout Rooms

In ELA 5/6, the facilitator started with a detailed discussion of each item after the first round of bookmark placements. Panelists provided content-based justifications and frequently referenced the threshold personas developed earlier in the week. “When I read that persona, and when I look at what that question is asking a student to do,” one participant said, “it is multiple skills being combined at one time.” Another added, “That’s what’s in the ALDs for the partially proficient.” These comments demonstrated the content-focused nature of the discussion, which continued as participants compared their item classifications with both the ALDs and their profiles of threshold students.

Across the math rooms, participants were similarly engaged in content-related discussions. In Math 3/4, the group discussed the OIB items, categorizing them into levels based on their alignment with ALDs and threshold descriptors. Math 5/6 participants explained their placements referencing both the KSAs and statistical indicators. One panelist described her indecision between two items that were two pages apart in the OIB, noting that she ultimately favored the simpler one because “the more math you ask them to do, the greater the chances that they’ll make a mistake.”

In Math 7/8, item-by-item discussions continued with explicit reference to ALDs. One participant stated, “I don’t think a novice could get this question,” and another noted that although an item didn’t explicitly ask for evaluating a square root, it did require that skill in order to solve the problem. The facilitator brought up the ALDs again to help the group determine the appropriate classification.

Similar discussions took place in ELA 3/4 and ELA 7/8. In ELA 3/4, panelists concentrated on whether a specific item required implicit or explicit information and used this distinction to

determine its level. A high p-value indicated that many students answered the item correctly, but panelists questioned whether this was due to student knowledge or the perceived ease of answering the item through eliminating wrong choices. One participant said, “We’re in a state with horses, so they would know how to answer this item,” citing cultural familiarity as a possible reason for the item’s low difficulty. The facilitator reminded participants, “When we’re thinking can they do it 67% of the time, that doesn’t mean they all get it right,” reinforcing the logic of the two-thirds probability standard for bookmark placements.

In ELA 7/8, the group discussed item classifications, reaching agreement on many, but also having extended debates where disagreements arose. The reasons given were mostly based on content, with frequent mentions of the ALDs and the group’s threshold descriptors. Text complexity was often cited. “I’m stuck on this because it’s a hard passage,” said one participant. “So if you get anything right on it, it’s advanced.” After finishing the OIB classifications, the group started Round 1 of their bookmark placements.

At around 10:30 AM, the Math 7/8 group finished Round 2 for grade 8 and moved on to a pre-vertical articulation discussion. (Participants had previously agreed that no third round was necessary for grade 8.) Participants reviewed their median bookmark placements and voted on possible adjustments. The facilitator led a structured conversation about each cut, asking them to vote on the number of pages they felt comfortable lowering if needed. For one cut, five panelists voted to move down one page, while three voted for two. One participant mentioned the easier vocabulary of the page below the median; another described the item there as “basic computation.” For another cut point, seven out of eight participants agreed they were comfortable with

adjustments up to four pages downward. Votes were also taken on moving cuts upward, with participants citing the ALDs when explaining their decisions. The facilitator recorded all votes.

Vertical Articulation Rooms

At 1:00 PM, Tim Walker led a session on vertical articulation. He thanked participants for their efforts and explained the goal of the next activity. A slide highlighted the guiding principle for vertical articulation: "Achievement levels from grade to grade should articulate with one another such that large changes in the percentage of student performances … do not show big differences from grade to grade." Participants were informed they would represent their grade or content group and were given various sources of information, including suggested threshold placements, panel-recommended ranges, 2025 impact data, and historical articulation data.

The conversation in the ELA vertical articulation room focused on the consistency of novice and advanced classifications across grades. One participant, speaking from a parent perspective, suggested tightening the advanced and novice categories, arguing that some grade levels had too many students in each. Another participant disagreed: "We kept saying err on the side of the kid." A colleague responded, "We chose not to do that, because then we're taking away opportunities for intervention if they need it." A third participant added, "I have kids that are doing well in math but not doing well on this test because of the reading," noting the impact of reading skills on math test performance.

Participants suggested raising the threshold for partially proficient, but a representative pointed out that the items immediately after the threshold were still classified as novice. Two adjustments were ultimately made to better align novice percentages across grades.

The ELA vertical articulation room mainly focused on refining the advanced cuts. The facilitator reminded participants, “It’s not like they need to be the same percentages across the grade levels.” Instead, the focus was on consistency in interpreting grade-to-grade trends, with grade 3 described as having more “beginning readers,” and grades 4 and 5 reflecting increasing text complexity. At 2:00 PM, participants reviewed the most difficult items in the OIB (items between pages 34 and 40). They discussed whether partially proficient students could reasonably be expected to answer some of them correctly. After the discussion, participants agreed that the advanced categories seemed appropriate. That was the last of my observations in the ELA vertical articulation room.

In the vertical articulation session back in Math, the discussion turned to the grade 5 proficient cut. Participants experimented with different placements and referred to their item maps. One noted that grade 5 is “always harder” across states.

Grade 5 panelists grappled with the fact that between Rounds 2 and 3, their interpretation of “would” had become stricter, causing an unusually early placement of the proficient bookmarks. Rania Rotou of New Meridian suggested that they set aside the outlier item and instead focus on the items beyond it. (A drawback of the Bookmark method is that panelists' judgments can be unduly influenced by difficult items appearing early in the OIB.) By the end of the day, participants completed reflection forms and were invited to record any justifications for their final placements, especially when those placements fell outside their committee’s originally recommended ranges.

Conclusion

The Montana July 2025 standard-setting workshop for the MAST assessment was conducted with high levels of participant engagement and professional facilitation. Across four days, panelists engaged in content-rich discussions, considered multiple sources of evidence, and demonstrated appropriate use of ALDs and threshold student profiles to inform their judgments. Facilitators responded effectively to participant questions and adapted their approaches when additional support or clarification was needed, particularly regarding the role of writing in ELA, and the need for a third round of judgments where warranted.

The inclusion of vertical articulation on the final day enabled participants to evaluate coherence across grade levels. Overall, participants offered content-based reasons for their judgments and suggested adjustments. Although some variation in approach and emphasis was seen across rooms, the process structure and participants' professionalism provided a solid foundation for defensible cut scores for MAST.

References

Lewis, D., Mitzel, H., & Green, D. (1996, June). Standard setting: A bookmark approach. In D. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D., Mitzel, H., Mercado, R., & Schulz E. (2012) The Bookmark standard setting procedure. In G. Cizek (Ed.), *Setting Performance Standards* 2nd Edition. New York, Routledge.

Mitzel, H., Lewis, D., Patz, R., & Green, D. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, ME: Erlbaum.

New Meridian. (2025, January). Montana Aligned to Standards Through-Year Assessments: Standard-Setting Plan: Grades 3–8 English Language Arts and Mathematics.

Observation of the Montana July 2025 MAST Standard Setting

This report was prepared for the Montana Office of Public Instruction (OPI) by Will Lorié, Ph.D., Senior Associate at the National Center for the Improvement of Educational Assessment (Center for Assessment).

Center for Assessment