OFFICE OF CHIEF OF RESEARCH AND DEVELOPMENT
Summary of State Board of Education Agenda Item
April 20, 2017

OFFICE OF STUDENT ASSESSMENT

06. Action: Recommended cut score on the Mississippi Assessment Program (MAP) 3rd Grade Reading Summative Assessment [Goals 1 and 2 – MBE Strategic Plan]

Background information: The standard setting validation process was completed on March 30, 2017. The validation process is designed to review and establish cut scores to meet the requirements of the *Mississippi Literacy – Based Promotion Act (*MLBPA*)*. The recommended standard passing scores were determined for performance levels 2 and 3. These new recommended cut scores apply to the reading and language portions of the MAP ELA assessment to meet the requirements of the MLBPA.

This item references Goals 1 and 2 of the *Mississippi Board of Education 2016-2020 Strategic Plan.*

Recommendation: Approval

Back-up material attached

Spring 2017 Standards Validation Plan for the Mississippi Assessment Program (MAP) 3rd Grade Reading Summative Assessment

Presented to the
Mississippi Department of Education
by Questar Assessment, Inc.

April 2017

5550 Upper 147th Street West
Apple Valley, MN 55124
(952) 997-2700
www.Questarai.com

Table of Contents

Purpose of the Document iv
Executive Summary v
1. Introduction 1
    1.1. Overview of the Document 1
    1.2. Overview of the Assessment 1
2. Standards Validation Overview 2
    2.1. Definition 2
    2.2. Materials 2
    2.3. Panelists 2
    2.4. Staffing 5
        2.4.1. Facilitator 5
        2.4.2. Other Staff 5
    2.5. Security 5
    2.6. Performance Level Descriptor (PLD) 5
    2.7. Ordered Item Booklet (OIB) 6
    2.8. Item P-Values 6
3. Ordered-Item Yes/No (OIYN) Method 7
    3.1. Comparison of Other Methods 7
    3.2. Rationale for Using OIYN 7
    3.3. OIYN Process 8
    3.4. Calculating the Cut Score 9
    3.5. Determining the 3rd Grade Reading Summative Assessment Scores that Link to ELA Assessment Scores 9
4. Standards Validation Process 10
    4.1. Tasks Completed Prior to the Meeting 10
    4.2. Tasks Completed During the Meeting 10
        4.2.1. Orientation and Training 11
        4.2.2. Review the Grade 3 ELA Operational Assessment 11
        4.2.3. Define the Borderline Student 11
        4.2.4. Practice Exercise 12
        4.2.5. Round 1 13
        4.2.6. Discuss Round 1 Results 14
        4.2.7. Round 2 15
        4.2.8. Discuss Round 2 Results 15
        4.2.9. Round 3 16
        4.2.10. Review Round 3 Results 16
        4.2.11. Evaluation Survey and Dismissal 16
5. Results 16
    5.1. Rounds 1, 2, and 3 16
    5.2. State Broad Review and Approval 20
6. Reliability and Validity Evidence 20
    6.1. Reliability Evidence 20

6.1.1. Intra-panelist Consistency 20
6.1.2. Inter-panelist Consistency 20
6.2. Evaluation Survey 21
6.3. Hambleton (2001) Validity Considerations 23
6.4. Procedural Validity Evidence 26
6.5. Internal Validity Evidence 26
6.6. External Validity Evidence 26
7. References 27
Appendix X: Agenda 28
Appendix X: Panelist Information 29
Appendix X: Performance Level Descriptor (PLD) 30
Appendix X: Borderline Student Worksheet 31
Appendix X: Item Rating Form 34
Appendix X: Orientation PowerPoint Presentation 35
Appendix X: Facilitator Script 36
Appendix X: Evaluation Survey 37
Appendix X: Evaluation Survey Results 38
Appendix X: Readiness Form 40

List of Tables

Table E.1. Final Cut Scores and Impact vi
Table 1.1. MAP 3rd Grade Reading Summative Assessment vs. Grade 3 ELA Assessment 1
Table 3.1. Expected Cut Scores and Panelist Variability for the MAP 3rd Grade Reading Summative Assessment using Grade 3 ELA Assessment as Reference Point 10
Table 5.1. OIYN Results for All Rounds 17
Table 5.2. Impact Data after Each Round 20
Table 6.1. Standard Deviations and Standard Errors 21
Table 6.2. Summary of Validity Evidence 23

List of Figures

Figure 2.1. Example of Ordered Item Booklet (OIB) 6
Figure 4.1. Illustration of the Borderline Student 12
Figure 5.1. Cut Score Distribution Plot with Markers by Table 17
Figure 5.2. Item-by-Item Ratings by Table 19
Figure 6.1. Mean Responses to Each Evaluation Area with Six Options 22

Purpose of the Document

This document presents the standards validation plan for the Mississippi Assessment Program (MAP) 3rd Grade Reading Summative Assessment. This document also presents a preview of the final report that will be created after the event is completed to allow the Mississippi Department of Education (MDE) and the Technical Advisory Committee (TAC) to provide any feedback well in advance of the report preparation.

Executive Summary

Mississippi educators will use an ordered-item yes/no (OIYN) standards validation method to review the estimated[1] Performance Level 2 (PL 2) and 3 (PL 3) cut scores for the Mississippi Assessment Program (MAP) 3rd Grade Reading Summative Assessment to determine if changes to the estimated cut scores are necessary. The meeting will take place in March 30, 2017 in Jackson, Mississippi. The standards validation is necessary because a new assessment administered for the first time in 2015–2016 (i.e., reading items from the MAP Grade 3 English Language Arts (ELA) assessment)[2] is being used to meet the requirements of the Mississippi Literacy-Based Promotion Act (SB2347), creating the need for a review of the PL 2 and PL 3 cut score.

A passing cut score will place students into one of two performance levels:

- Not Pass (i.e., does not meet the readiness requirement)
- Pass (i.e., meets the readiness requirement)

However, two passing standards will be determined. The first, based on PL 2, will be the initial passing cut score for the MAP 3rd Grade Reading Summative Assessment. The second cut score, based on PL 3, is being established in the event that the Mississippi Department of Education (MDE) wishes to phase in a more rigorous passing standard at a later date.

The MDE will recruit panelists who have taught reading at Grades 3 and 4 and are experienced with the Mississippi College and Career Readiness Standards (MS-CCRS). A total of 12 participants will be selected: eight Grade 3 educators and four Grade 4 educators. Panelists will be divided across two tables in order to provide an indication of variability across distinct groups of panelists.

The standards validation event will include the following activities:

1. Orientation, training, and practice
2. Discussions about the knowledge and skills possessed by borderline PL 2 and PL 3 students
3. Three rounds of the OIYN procedure for each grade
4. State Board review and approval

Table 1.1 presents a preview of how the final raw, theta, and scale score cut scores that are approved by the State Board of Education will be presented. It will also present the percentage of students who took the reading operational items from the MAP Grade 3

[1] Estimated Gateway cut scores were determined by linking the Gateway Assessment items to the Grade 3 ELA assessment.
[2] The Grade 3 Gateway Assessment used the last two years was developed by Renaissance Learning. The Renaissance Learning test will continue to be used as retest for students who do not past during the first assessment.

ELA assessment[3] in Spring 2016 who fell into each performance level based on those cut scores.

Table E.1. Final Cut Scores and Impact

	Cut Score			Impact	
Cut Level	**Raw**	**Theta**	**Scale**	**%Not Passing**	**%Passing**
PL 2 Cut Score					
PL 3 Cut Score					

***Comment*. Table will be completed after the event.**

Notes. The PL 3 cut score is available in the event that the Mississippi Department of Education (MDE) wishes to phase in a more rigorous passing standard at a later date.

[3] The Gateway assessment only includes reading items from the MAP Grade 3 ELA assessment.

1. Introduction

1.1. Overview of the Document

This document presents the standards validation plan from Questar Assessment, Inc. (Questar) for the Mississippi Assessment Program (MAP) 3rd Grade Reading Summative Assessment. This standards validation is necessary because a new assessment (i.e., reading items from the MAP Grade 3 English Language Arts (ELA) assessment) is being used to meet the requirements of the Mississippi Literacy-Based Promotion Act (SB2347), thus creating the need for a passing cut score, or the minimum score a student must achieve on the assessment in order to pass.

During the one-day standards validation meeting in Spring 2017, two panels of Mississippi educators will follow an ordered-item yes/no (OIYN) standards validation method to review two estimated passing cut scores, one based on PL 2 and one PL 3, that distinguish the Not Pass and Pass performance levels for Grades 3 students to determine if changes are necessary. They will use the performance level descriptors (PLDs) and items from the Spring 2016 test forms, ordered by difficulty based on student performance on the Spring 2016 test.

The resulting recommended cut scores will then be provided to the Mississippi Department of Education (MDE) to present for final approval from the State Board of Education.

1.2. Overview of the Assessment

The 3rd Grade Reading Summative Assessment meets the requirements of the Mississippi Literacy-Based Promotion Act (SB2347). This bill requires Grade 3 students to demonstrate basic level reading proficiency in order to be promoted to Grade 4. The 3rd Grade Reading Summative Assessment is aligned to the Mississippi College and Career Readiness Standards (MS-CCRS). The newest MAP Assessments are based on these standards and were first administered in the 2015–2016 school year.

As indicated in Table 1.1, the 3rd Grade Reading Summative Assessment will consist exclusively of the operational reading items from the MAP Grade 3 ELA Assessment (reading literature and reading for information). The maximum possible 3rd Grade Reading Summative Assessment score is 40 points, which makes it worth fewer points than the MAP Grade 3 ELA Assessment because the ELA test also includes language and writing tasks. As the reading items make up two-thirds of the ELA score, there will be a strong association between the ELA and Reading assessment scores. Currently, reading items only include multiple-choice, M of N, and drag-and-drop item types.

Table 1.1. MAP 3rd Grade Reading Summative Assessment vs. Grade 3 ELA Assessment

	Test Blueprints*				Mean Difficulty		SD Difficulty	
Test	RL	RI	L	W	*p*-value	Logit	*p*-value	Logit
ELA	20	20	8	12	0.488	0.000	0.195	1.077
Reading	20	20	0	0	0.462	0.062	0.196	1.081

*RL= Reading Literature. RI = Reading for Information. L = Language. W = Writing. Differences in content are underlined.

2. Standards Validation Overview

This section describes the definition of standard's validation and the general components of the standards validation meeting, including the list of materials, panel composition, staffing, security, the PLD, the ordered item booklet (OIB), and the item map (dot plot). The following sections then describe the OIYN in detail and chronicle the standard setting steps that occurred during the meeting.

2.1. Definition

Standards validation is a formal process by which committees of educators and subject matter experts reconsider performance standards, or cut scores, when there is a minor change in a testing program, such as the addition of new item types. In a standards validation, some additional scaffolding is available for panelists (e.g., a recommended starting point is provided for their consideration). The end result is the same as a traditional standard setting: cut score(s) divide a score scale into performance levels (e.g., Not Pass and Pass) that students are placed into based on their test results.

2.2. Materials

The following materials will be used during the standards validation meeting:

- Agenda (Appendix **X**)
- Panelist information form
- Nondisclosure agreement (Appendix **X**)
- Reimbursement form (Appendix **X**)
- Orientation and training PowerPoint presentation (Appendix **X**)
- 2015–2016 test items
- MS-CCRS
- PLD (Appendix **X**)
- Borderline student worksheet
- Practice passage and items
- Practice item rating form
- Readiness form (Appendix **X**)
- Ordered item booklets (OIBs)
- Item rating form (Appendix **X**)
- Impact data
- Evaluation survey (Appendix **X**)
- Facilitator script (Appendix **X**)

2.3. Panelists

As shown in Table 2.2, twelve participants will be recruited by the MDE. Table 2.3 then presents a preview of how the panelists' demographic characteristics will be presented in the final report. This information will be collected in a panelist information form provided in Appendix **X**, which will be used to collect panelists' background information,

including gender, ethnicity, current work assignment and setting, teaching experience, and familiarity with the MS-CCRS. This information will be collected to determine if the panels comprised a fair and representative sample of the state's educators. Names will not be shown for privacy purpose.

Panelists will have taught reading at Grades 3 and 4 and will be experienced with the MS-CCRS. Grades 3 and 4 will be separated into two tables (each with four Grade 3 teachers and two Grade 4 teachers) in order to allow for more in-depth discussion among panelists. The selection and training of the standards validation panelists will be crucial to the success of the meeting. During the selection of the panelists, the MDE will consider several aspects of panel diversity, including gender, ethnicity, geographic location, and teaching experience).

Table 2.2. Panel Composition

Table	#Panelists
Grade 3	8
Grade 4	4
Total	**12**

Table 2.3. Panel Demographics

	Freq.	%
Gender		
Female		
Male		
Ethnicity		
White		
Black		
Other		
Role		
Classroom Teacher		
Nonteacher Educator		
Other		
Region		
Urban		
Suburban		
Rural		
MS-CCRS		
Familiar		
Not familiar		
#Years Experience		
< 5 years		
5–10 years		
> 10 years		

***Note*. Table will be completed after the event.**

2.4. Staffing

2.4.1. Facilitator

A psychometrician experienced in facilitating standards validation events will facilitate the panelist group. The facilitator will serve the following functions:

- Guide the panelists through the standards validation process.
- Provide feedback and answer questions.
- Analyze data at the end of each round to prepare for the next round.
- Provide feedback data and facilitate discussions between rounds.

A script (see Appendix X) will be prepared ahead of time for the facilitator. The script will cover all major procedural elements used at the standards validation. The script is intended to:

1. Make sure the process is standardized to minimize any potential facilitator artifacts
2. Help meet all timelines

The script will only suggest a basic narrative for explaining procedures. The facilitator does not need to read the script verbatim as the long key elements are conveyed to panelists. The sequence of procedures will be standardized and cannot be altered.

2.4.2. Other Staff

If needed, an additional Questar psychometrician or statistical analyst will assist with data analysis, oversee data quality control, and observe the activities. Questar assessment specialists will also be present to address content questions. Additional Questar staff will be available for addressing other administrative tasks, and MDE staff can observe any or all parts of the standards validation.

2.5. Security

Printed materials will be needed during the standards validation meeting. Panelists will be required to leave all materials in their room during breaks and at the end of the meeting. Questar will monitor materials throughout the day, including lunchtime. Additionally, panelists will sign non-disclosure agreements before beginning the standards validation judgments. Facilitators will continuously remind panelists about the security policies throughout the meeting, emphasizing that the security of testing materials should be maintained at all times.

2.6. Performance Level Descriptor (PLD)

The performance level descriptor (PLD) describes the set of knowledge, skills, and abilities that students are expected to display in order to pass the 3rd Grade Reading Summative Assessment. The 3rd Grade Reading Summative Assessment PLD[4], which has already been developed and is provided in Appendix X, is a narrative descriptor of

[4] This PLD is a subset of the 3rd Grade ELA Assessment involving reading standards (no language or writing)

the performance level linked to the MS-CCRS. The PLD contains important information for panelists to consider when making their item-level judgments.

2.7. Ordered Item Booklet (OIB)

Test items will be presented to panelists in OIB in order of item difficulty, going from the easiest item on the first page to the hardest item on the last page, as shown in Figure 2.1.

For the 3rd Grade Reading Summative Assessment items, Rasch calibration values (item difficulty indices) have already been computed based on the 2016 operational Grade 3 ELA test data. For each score point (except zero-point) of a polytomous item, cumulative Rasch probability curves indicating the probability of a student scoring at and above at a given score point will be calculated to reflect the difficulty of earning each score point or any greater. All two-point items in the 3rd Grade Reading Summative Assessment will have two Rasch values corresponding to scores of one (or greater) and two. Based on these Rasch v alues, each one-point item will appear once in the OIB, while each two-point item will appear twice, one for each score point greater than zero, as shown in Figure 2.1. This is necessary because each score point is associated with a different difficulty level.

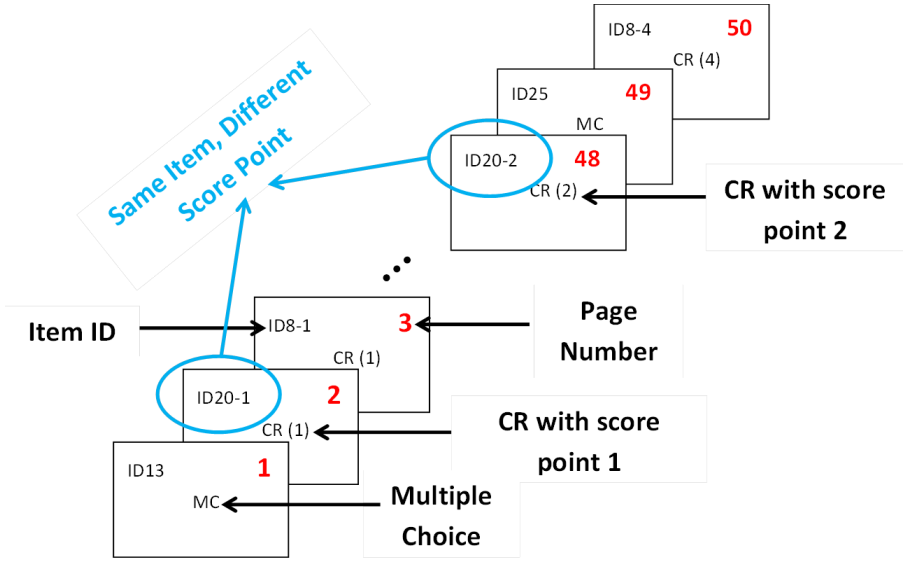


Figure 2.1. Example of Ordered Item Booklet (OIB)

2.8. Item P-Values

A primary purpose of the OIB is to reduce the cognitive complexity of the panelists' item judgments by providing a relative indication of the item difficulties via the ordering of the items. However, the OIB only provides information about the relative item difficulties. Panelists can benefit from more absolute estimates of item difficulty as they make their yes/no decisions. To foster more accurate panelists' judgments, the empirical difficulties of items in the form of item *p*-values (the percentage of students who got an item correct) will be provided to panelists.

3. Ordered-Item Yes/No (OIYN) Method

3.1. Comparison of Other Methods

The ordered-item yes/no (OIYN) method combines advantages of the Angoff and Bookmark methods (summarized below) while also addressing some of the issues related to each method.

- The traditional Angoff (1971) standard setting method requires panelists to review each test item and estimate what proportion of a hypothetical group of minimally competent examinees (i.e., borderline examinees) would answer each item correctly. Angoff panelists may have difficulty in (1) conceptualizing the hypothetical borderline examinee, or (2) estimating the proportion correct on the items for the borderline students (i.e., conditional *p*-values). These challenges can affect panelists' judgments about cut scores which in turn can contaminate the validity of the cut scores (Hambleton & Pitoniak, 2006; Shepard, Glaser, Linn, & Bohrnstedt, 1993; Smith & Davis, 2009).

- In contrast, the Angoff yes/no method (Impara & Plake, 1997) has panelists consider an actual examinee believed to be on the borderline who is known to them, and then make a simple dichotomous (yes or no) judgment about whether their prototypical borderline examinee would be able to answer each question correctly. Since the Angoff yes/no method simplifies the judgmental task, its use is expected to be clearer to panelists and hence easier to use than the Angoff probability-estimation procedure (Impara & Plake, 1997).

- In the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001), items are presented in an ordered item booklet (OIB) from easiest to hardest based on empirical item difficulty estimates. Panelists review items in order and place a bookmark at the page in their OIB at the point where they believe the borderline examinee would not have a certain probability (e.g., 67%) of answering the item correctly. Panelists must take the item ordering as a given when they consider their bookmark placements. However, Bookmark facilitators have noted that panelists sometimes think that some harder items (i.e., those appearing later in the OIB) should be easier than some items appearing earlier in the OIB, and vice versa (Smith & Davis, 2009). Such dissonance can affect panelists' bookmark placements. For example, Skagggs and Tessema (2001) found that panelists who questioned the item ordering in the OIB had higher bookmark placements than other panelists.

Smith and Davis (2009) studied an OIYN method, which takes advantage of the OIB scaffolding from the Bookmark method, reducing the complexity of the panelists' task of judging item difficulties since the items are ordered by difficulty level. The method follows Angoff procedures by asking panelists to make multiple item-level judgments rather than locate a single Bookmark position. In this way, panelists make individual decisions about each item.

3.2. Rationale for Using OIYN

The 3rd Grade Reading Summative Assessment has fewer points than the Grade 3 ELA assessment, which may have led to the significant gaps between the item difficulties observed in a preliminary OIB constructed to evaluating the item pool (visualized in Figure 3.2). This, in turn, would lead to difficulty in determining where exactly to set the cut scores with the traditional Bookmark method.[5]

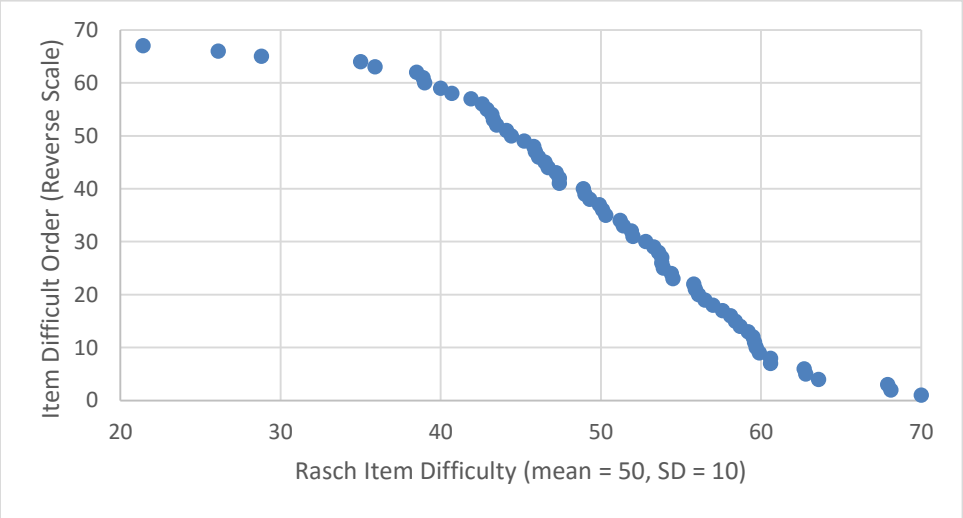


Figure 3.1. Dot Plot of Difficulties for All Reading Items in the Grade 3 ELA Bank

3.3. OIYN Process

During the standards validation meeting using the OIYN method, panelists will review each item in the OIB during three rounds of rating. Panelists will be instructed to think about whether the borderline PL 2 and PL 3 student would answer each item correctly. Only a dichotomous "Yes" or "No" response will be required for each item. For Rounds 1 and 2, panelists will be asked to indicate their yes/no judgments for each item on the item rating form (presented in Appendix X). For Round 3 (the final round), panelists will only need to write down the minimum number of points required to pass the 3rd Grade Reading Summative Assessment.

Since the items are rank-ordered in difficulty from the easiest to hardest, panelists are expected to make more "Yes" judgments at the beginning of the OIB (easier items) and more "No" judgments at the end of OIB (more difficult items). Ideally, each panelist will come to a point where all the "Yes" answers would change to "No" answers. Panelists will be told that the item ordering provides useful information, but is not an absolute rule for answering "Yes" or "No."

[5] Consider a median Bookmark placement falling between two items with a large theta gap: 1.00 and 1.50. If more items were available within this range, the Bookmark may have fallen closer to 1.00 or 1.50.

3.4. Calculating the Cut Score

In order to determine the cut scores, the panelists' "Yes" responses over all items[6] will be converted to ones (1's), and their "No" responses will be converted to zeros (0's). The total number of 1's will be considered the cut score on the raw score scale for each panelist.[7] The median cut score across all panelists will be determined and converted to an ability (theta) value, representing the panelists' estimated theta cut for the passing score.[8]

3.5. Determining the 3rd Grade Reading Summative Assessment Scores that Link to ELA Assessment Scores

In 2015, roughly 85% of first-time test takers passed the 3rd Grade Reading Summative Assessment on their first attempt. In 2016, this increased to about 87%. Also in 2016, just over 90% of the students scored at the Basic performance level or higher on the MAP Grade 3 ELA Assessment.

A linking study was undertaken to determine what scores on the reading items from the 2016 MAP Grade 3 ELA Assessment correspond to different cut scores on the full 2016 MAP Grade 3 ELA Assessment. That linked PL 2 and PL 3 scores will be used to identify a range of items for the panelists to focus on during the standard's validation.

Specifically, once the linked PL 2 and PL 3 scores are determined, a score interval of +/- one standard errors around the PL 2 score will be determined (i.e., PL 2 score – 1 * SE, PL 2 score + 1 * SE). Any item below the lower threshold will be considered an automatic "Yes" response[9], and any item above the upper threshold will be considered a "No" response. In other words, the panelist will be asked to focus their attention on the items between the lower and upper thresholds.

Table 3.1 provides information about variability in student test scores (e.g., SEMs) as well as expected variability in panelists item ratings (based on median absolute deviation—MAD from the Grade 3 ELA standard setting conducted in Spring of 2016). The expected raw score cuts are 11 and 17, respectively. The largest MAD from the last standard setting was 4, but that was for the longer ELA test. The 3rd Grade Reading Summative Assessment CTT SEM is about 3. Intervals of 11 ± 9 and 17 ± 9 seem reasonable and conservative item zones for panelist to focus on during the standard setting.[10]

[6] Or in the case of polytomous items, each score point.
[7] A correction for guessing could be considered in the cut score calculation, but it is not recommended. A guessing correction is not consistent with the item scoring (where no guessing correction is applied) or scaling (where no IRT guessing parameter is estimated). Further, a higher cut score could result, which would not give the benefit of the doubt to students.
[8] Other methods could be considered (e.g., logistic regression). However, it is not clear if other methods offer a clear advantage in the context of this study.
[9] For simplicity panelist may start with Item 1 in the OIB even if it is below the lower threshold.
[10] Later, an ordered item booklet (OIB) will be introduced. Having panelist focus on items 1 to 26 in the OIB would cover both zones of interest.

Table 3.1. Expected Cut Scores and Panelist Variability for the MAP 3rd Grade Reading Summative Assessment using Grade 3 ELA Assessment as Reference Point

	PL2		PL3		PL4		PL5	
	ELA	**Reading**	**ELA**	**Reading**	**ELA**	**Reading**	**ELA**	**Reading**
Scale Score	335	335	350	350	365	365	387	387
IRT CSEM	6	7	6	7	6	7	7	8
Raw Score Cut	18	11	27	17	36	23	47	31
CTT SEM	3.4	2.8	3.4	2.8	3.4	2.8	3.4	2.8
MAD	1	--	4	--	1	--	2	--

Note. The median absolute deviation (MAD) indicates the variability in panelist raw score cuts for the Grade 3 ELA standard setting conducted in Spring of 2016.

4. Standards Validation Process

4.1. Tasks Completed Prior to the Meeting

Tasks to be completed prior to the standards validation meeting include the following:

- Creation of meeting agenda
- Selection of panelists
- Preparation of test booklet containing operational reading items from the Grade 3 ELA assessment
- Creation of the PLD
- Preparation of the borderline student worksheet
- Creation of OIB
- Preparation of item map (dot plot)
- Preparation of training materials
- Preparation of facilitator script
- Preparation of item rating form
- Creation of evaluation survey
- Preparation of other administrative forms (e.g., nondisclosure form)
- Development of software for data entry and analysis

4.2. Tasks Completed During the Meeting

Appendix X presents the meeting agenda. During the standards validation meeting, the following activities will take place:

- Orientation and Training
- Review the Grade 3 ELA Operational Assessment
- Define the Borderline Students
- Practice Exercise
- Round 1
- Discuss Round 1 Results
- Round 2

- Discuss Round 2 Results
- Round 3
- Review Round 3 Results
- Evaluation Survey

4.2.1. Orientation and Training

Appendix **X** contains the orientation and training PowerPoint presentation. The MDE will begin the standards validation event by welcoming the panelists and establishing the importance of their work. Next, the facilitator will introduce the purpose and goal of the standards validation and the roles and responsibilities of those involved in the event.

The facilitator will then give a step-by-step overview of the OIYN method. The objective, the task to be completed, and the materials to be used will be described for each step in the process. Important concepts such as the borderline student and how to use the item rating form will be emphasized. Panelists will be encouraged to ask any questions they might have about the method, procedures, and documents. At the end of the orientation, administrative issues will be addressed such as event security.

The orientation will also provide historic passing rates for the Renaissance Learning test to give panelists a benchmark as they review the current cut score and determine if changes are necessary.

Next, the panelists (including the facilitator) will introduce themselves in an icebreaker. The facilitator will then stress the confidentiality of the items and have the panelists sign a nondisclosure form.

4.2.2. Review the Grade 3 ELA Operational Assessment

The first formal task will be to review the operational reading items from the Grade 3 ELA Assessment. Each panelist will receive a paper copy of the operational reading items ordered by their test administration sequence. Even though the MAP assessments are generally administered on the computer, this is not considered a major threat to the validity of the study because the reading item types are selected-response in nature (i.e., multiple-choice, M of N, and drag-and-drop).

The purpose of this activity will be to give panelists an understanding of the test content and the item and test difficulty. The facilitator will distribute the operational items to the panelists, who will spend about 30 minutes answering the items. During this process, panelists will consider the items and test from the students' perspective and think about the kinds of knowledge and skills measured by each item. When all panelists have finished the test, they will engage in a short discussion about the test and items.

4.2.3. Define the Borderline Student

Before making their yes/no judgments, panelists will review the PLDs and develop a common understanding of the specific knowledge and skills that borderline students between Not Pass and Pass have. Questar will prepare two borderline student worksheets for this process—one corresponding to the PL 2 ELA performance level and one corresponding to the PL 3 ELA performance level (see Appendix **X** for an example).

The PLD statements will be listed in the left-side column of the worksheet. In the right-side column, panelists will write down their list of specific attributes for the borderline student. Panelists' notes related to reading skills for *borderline PL 2* and *PL 3* students from the Grade 3 ELA standard setting will be included in the borderline student worksheet as well.

The purpose of this step is to help the panelists achieve a common understanding of the knowledge, skills, and abilities required to be classified as Passing. The image in Figure 4.1 will be provided to help panelists visualize the concept of the borderline student.

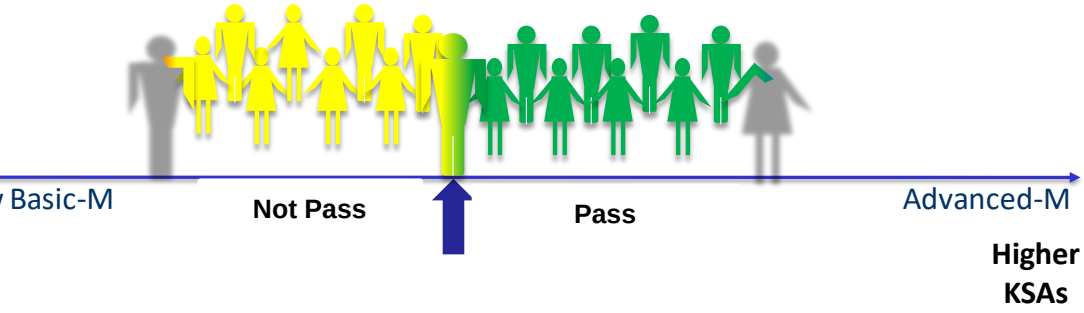


Figure 4.1. Illustration of the Borderline Student

The facilitator will emphasize that this step is very important since the focus of the standards validation is on the borderline student. The facilitator will also tell panelists that their task is not to edit the PLD. Instead, it is to use the PLD to further drill in on the content knowledge and skills that the borderline student has.

For the borderline student, panelists are to consider "what would demonstrate just enough knowledge or skill level to Pass?" based on the PLD statements. They will be reminded that in order to make each listed PLD statement concrete regarding the borderline students, they should consider behaviors and classroom experiences directly linked to the content standards and focus on content knowledge, skills, and abilities (not student attributes like social economic status). The small group discussions will take about 30 minutes. Panelists will be asked to develop at least one borderline attribute for each PLD statement, which they will write on their worksheets.

Once the small group discussion is over, panelists will discuss the brainstormed ideas in a large group. They will work on each PLD statement one at a time. The large group discussion will also take about 30 minutes. Finally, the consensus borderline student attributes will be printed out for the panelists to use in Rounds 1, 2, and 3.

4.2.4. Practice Exercise

Before starting Round 1, the panelists will be given about 30 minutes to practice making yes/no item decisions. The facilitator will project the first practice item for all panelists to see and ask them to use the PLD and their just developed borderline student worksheets to answer the question, "Would the borderline PL 2[11] student be able to

[11] Or PL 3

answer this item correctly?" For the practice polytomous item, panelists will answer the question, "Would the borderline PL 2[12] student be able to earn this score point or any higher?"

After answering, panelists will share their thoughts and rationales for their answers. Next, the panelists will consider the second practice item and answer the same question. Panelist discussion for the second item will then occur. This process will continue for all practice items.

4.2.5. Round 1

After the practice session, the facilitator will explain the task for Round 1. Panelists will individually review each item in the OIB (within the range of items identified by Questar) and determine whether a borderline PL 2 and PL 3 student would answer the item correctly (or would earn a particular score point or any higher for polytomous items). The key objective for Round 1 is to obtain the panelists' preliminary cut scores only using the PLD and borderline student worksheet as guides.

After the facilitator has reviewed the Round 1 task, he or she will distribute the readiness form and ask all panelists to fill it out. Panelists' responses to this form will provide evidence about their understanding of the event procedures, their knowledge about how to use different information to guide their ratings, and their preparedness to do their ratings.

When all panelists have given an affirmative answer on the readiness form, they will make their Round 1 yes/no decisions individually using the OIB, PLD, borderline student worksheet, and item rating form. The facilitator will emphasize that the panelists have to work alone and that no group discussion is permitted. This will help achieve independence among the Round 1 ratings.

The panelists will start with the first item identified by Questar in the OIB (most likely item 1). Just as they did with the practice items, panelists will be asked to consider both the PLD and the borderline student worksheets to answer "Yes" or "No" regarding whether a borderline PL 2 or PL 3 student would answer the first item correctly. After panelists make their decisions for the first item, they will record it on the item rating form. They will then proceed to the next item and do the same thing. This will continue until the panelists have recorded answers to all the items for the borderline PL 2 and PL 3 student on their item rating forms.[13]

Panelists will be reminded that items are ordered by difficulty in the OIB and that, because of this, they should have a consistent string of "Yes's" for the earlier items in the OIB and a consistent string of "No's" for the later items in the OIB. However, this will be described as a strong recommendation versus an absolute rule. A panelist may provide a disordered yes/no answer, although he or she must have a strong rationale recorded on the back of their item rating form.

[12] Or PL 3

[13] Panelists will start by answering questions for the Basic level first and then do the Passing level next.

Panelists will sum their number of "Yes" answers and record the result at the bottom of their item rating form. The facilitator will monitor the room to make sure that all panelists use the item rating form correctly.

4.2.6. Discuss Round 1 Results

4.2.6.1 Calculate the Results

When Round 1 is complete, the facilitator will collect the item rating forms and the panelists will take a short break.[14] Questar staff will enter the Round 1 data into Excel and execute a program that will be developed to carry out all the analyses during the standards validation event. Using this program, a summary of their ratings and the resulting impact data will be calculated upon the program's execution. This program will also provide graphical displays of the panelists' ratings. The program will be thoroughly tested prior to the event.

When entering the Round 1 results, Questar staff will first double-check that the panelists' item rating forms have no errors. Summary tables including the minimum, median, and maximum cut scores (i.e., sum of "Yes's") will be created. Two summary charts will be also created to capture information about the variations in the panelists' ratings. A relative frequency graph will show the number of panelists with a given number of "Yes's." The other chart will show the frequency of "Yes" answers for each OIB item.

4.2.6.2 Review the Results

Panelists will then review the Round 1 results. Only table-level feedback will be given at this time. Results for the complete panel will not be shown until after Round 2.

The facilitator will first show the Round 1 summary tables (i.e., the minimum, maximum, and median number of "Yes's" per table) and the individual panelists' number of "Yes" answers. A chart will show the frequency of panelists' personal cut scores. Panelists will be asked to compare their own results with those of others at their table and think about (1) how strict or lenient they are in relation to others and (2) how their individual experiences and perspectives may have affected their expectations for student performance. The panelists who have fewer "Yes's" or more "Yes's" will be allowed to share their rationales with the group.

A chart of the frequency of "Yes" answers on each item will be presented. For this figure, the panelists will be asked to give additional consideration to any items where the panelists' answers are inconsistent (i.e., items were some panelists answered "Yes" but other panelists answered "No"). They will also think about any factors that may have caused inconsistencies in their answers. During this phase, panelists will be encouraged to share their thoughts with others at their table.

4.2.6.3 Review New Information

[14] Panelists will leave all secure materials in the room during their breaks.

After the group discussion, panelists will be given new information. First, the facilitator will show a dot plot of the overall item difficulties and explain how to interpret the figure. Panelists will review the items where they disagreed, again in consideration of the specific item-difficulty information. Panelists will be allowed to share their thoughts with the large group.

Impact information (i.e., the percentages of students passing and not passing based on PL 2 and PL 3) will be given, along with the impact observed in past administrations. The percentages will be based on the students who actually took the reading items from Spring 2016 ELA test. Once again, panelists will be allowed to share their thoughts with the large group.

4.2.7. Round 2

After the facilitator describes the task for Round 2, panelists will reconsider their Round 1 answers while integrating information about the Round 1 results, the item difficulty dot plot, the impact data, and the group discussions. Panelists can adjust any of their Round 1 answers if they want to. The facilitator will emphasize that the fundamental question remains the same: “Would the borderline student be able to answer the item correctly, Yes or No?” They will be reminded to still consider the PLD and the borderline student attributes.

Panelists will not need to conform to their Round 1 median rating, although they should consider that piece of information. Panelists will also still need to make their judgments individually.

When making their Round 2 ratings, the panelists are to:

- Proceed sequentially through the OIB
- Document their answers on the item rating form
- Sum up their “Yes” answers

After Round 2, the panelists will be dismissed for a break. Questar staff will then enter the Round 2 ratings and analyze the results. The summary tables and charts prepared for Round 1 will be also prepared for Round 2. Updated impact information will also be calculated.

4.2.8. Discuss Round 2 Results

The statistical results for Round 2 will be provided for the entire panel and broken out by tables. The facilitator will again show panelists their individual numbers of “Yes” answers; the minimum, maximum, and median number of “Yes’s;” and the summary charts showing the variation in their numbers of “Yes’s.” As before, panelists will compare their ratings with others and consider if they are strict or lenient. Panelists will share their thoughts and rationales with the group.

Next, the facilitator will show the updated impact data and communicate that this represents the updated percentages in the two performance levels if the cut scores were based on the median of the two table’s median number of “Yes” answers from

Round 2. The panelists will think about the updated percentages and discuss their thoughts in a large group.

4.2.9. Round 3

The facilitator will introduce the Round 3 task, emphasizing that Round 3 will be the final opportunity for the panelists to revise their prior results. In other words, Round 3 will provide the "provisional" cut score that will be taken to the State Board of Education.

The Round 3 ratings will be conducted slightly differently from the previous two rounds. Specifically, the panelists will not provide yes/no answers item by item. Instead, they will only need to provide one number: the minimum number of points needed to Pass. The panelists will use all the available information to help guide their final decision (e.g., the results from the prior two rounds, the impact data, the discussions with their colleagues, the PLD, and the borderline student attributes). As before, panelists will independently record their final cut scores on the item rating form.

After Round 3, the panelists will be dismissed for a break. Questar staff will enter their final ratings and produce the tables and charts prepared for prior rounds (e.g., the minimum, maximum, and median numbers of points needed to pass; summary charts of the panelists' cut scores; and updated impact data based on the median of the two tables' median cut score values).

4.2.10. Review Round 3 Results

After the break, panelists will come together to review the final round's results. Panelists will be encouraged to share their final thoughts. Once again, the facilitator will remind panelists that these result will be provisional cut scores to be reviewed by the State Board of Education.

4.2.11. Evaluation Survey and Dismissal

An evaluation survey will be given to panelists after Round 3. The survey questions will cover several dimensions, including panelists' opinions on the overall process and, perhaps most importantly, their confidence in their final recommended cut score. The facilitator will collect all the secure materials and the completed evaluation surveys. On behalf of the state, panelists will be thanked for their hard work.

5. Results

A program will be developed to carry out all the analyses during the standards validation event. Staff will input the panelists' yes/no ratings in an Excel spreadsheet. A summary of their ratings and the resulting impact data will be calculated upon the program's execution. This program will also provide graphical displays of the panelists' ratings. The program will be thoroughly tested prior to the event.

5.1. Rounds 1, 2, and 3

Table 7.1 provides an example of how the panelists individual cut scores over all three rounds will be presented. For Rounds 1 and 2, this corresponds to the number of "Yes" answers on the item rating form. Overall summary statistics (minimum, median, and

maximum values) will be provided as are disaggregated results by table. The overall median will be the median number of "Yes's" over the two tables, which will be used as the basis for determining the recommended raw score cuts at each round.

Table 5.1. OIYN Results for All Rounds

	PL 2 Cut Score			PL 3 Cut Score		
Panelist	**Round 1**	**Round 2**	**Round 3**	**Round 1**	**Round 2**	**Round 3**
Table 1						
1						
2						
3						
4						
5						
6						
Minimum						
Median						
Maximum						
Table 2						
7						
8						
9						
10						
11						
12						
Minimum						
Median						
Maximum						
Overall						
Minimum						
Median						
Maximum						

***Note*. Table will be completed after the event.**

Figures 5.1 presents an example of the graphical illustrations that will be presented to panelists during the meeting to visualize the variability in the panelists' cut scores for Rounds 1, 2, and 3, respectively. In these graphs, the *x*-axis is the raw score scale that ranges from 0 to 40. Each dot represents a panelist and is color coded by table membership. The red vertical line in each graph represents the overall median cut scores.

Figure 5.1. Cut Score Distribution Plot with Markers by Table

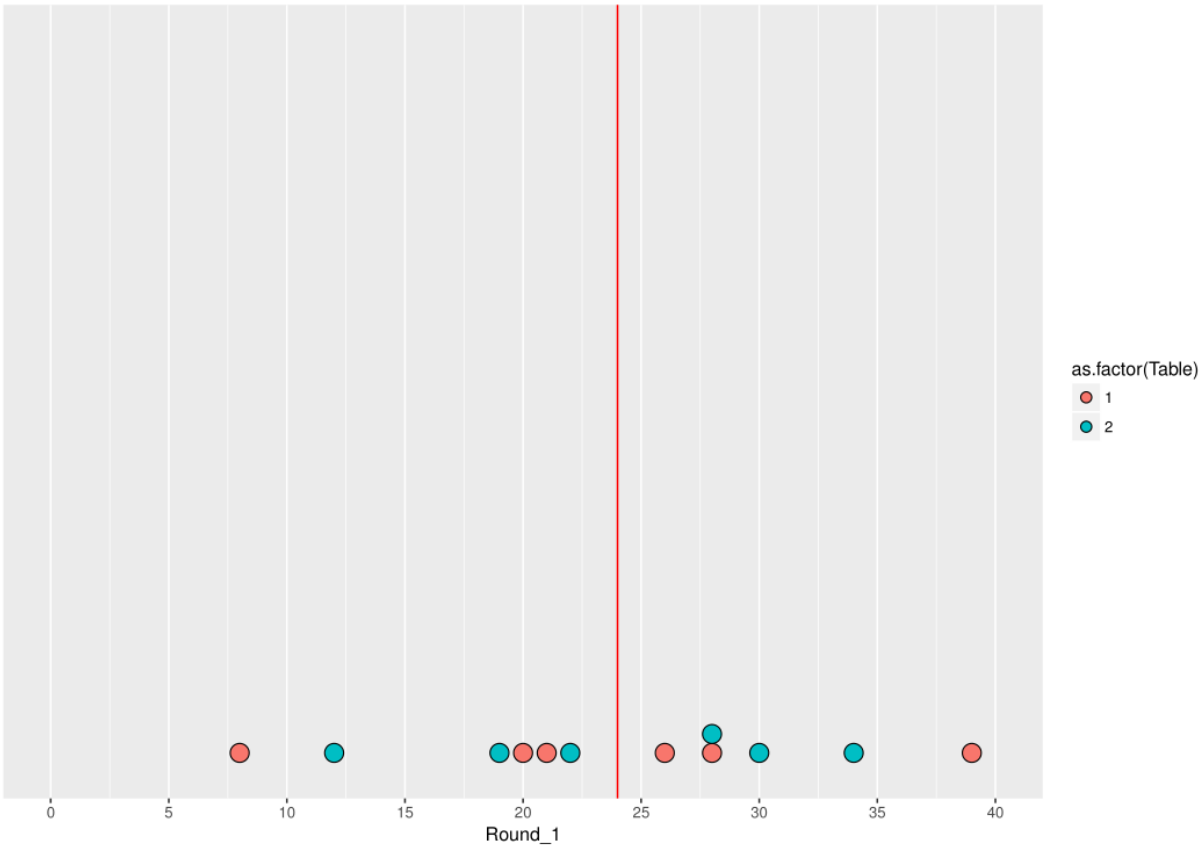


Figure 7.4 provides another example of a visualization reflecting the variability in the panelists' responses to individual test items.

Figure 5.2. Item-by-Item Ratings by Table

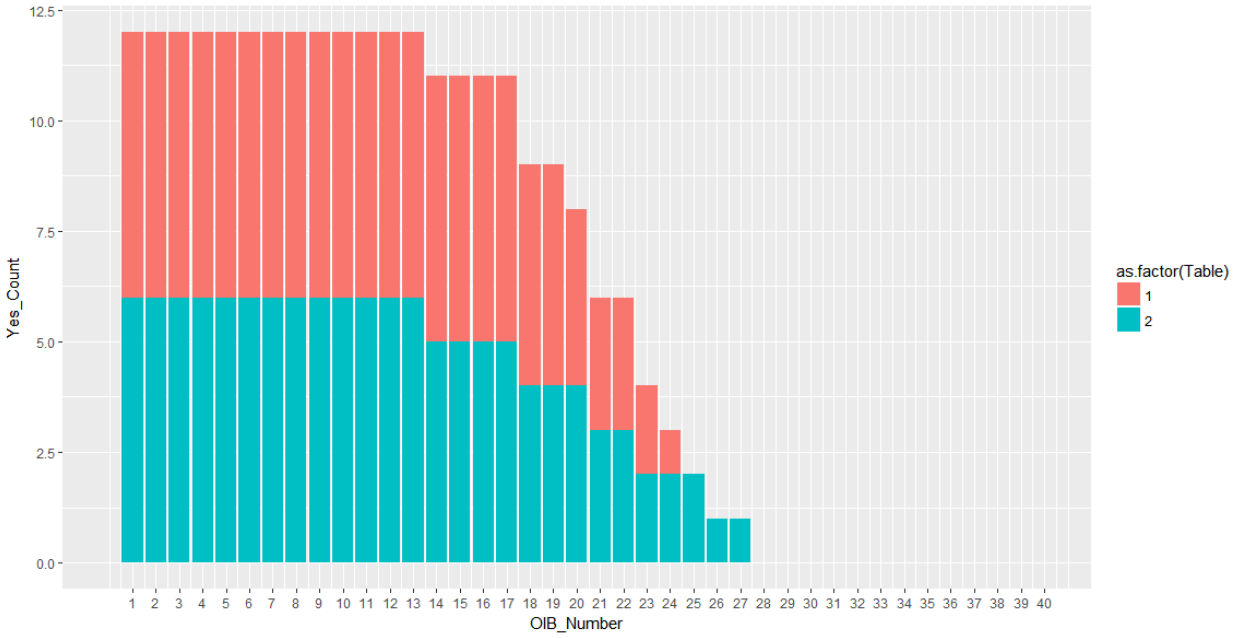
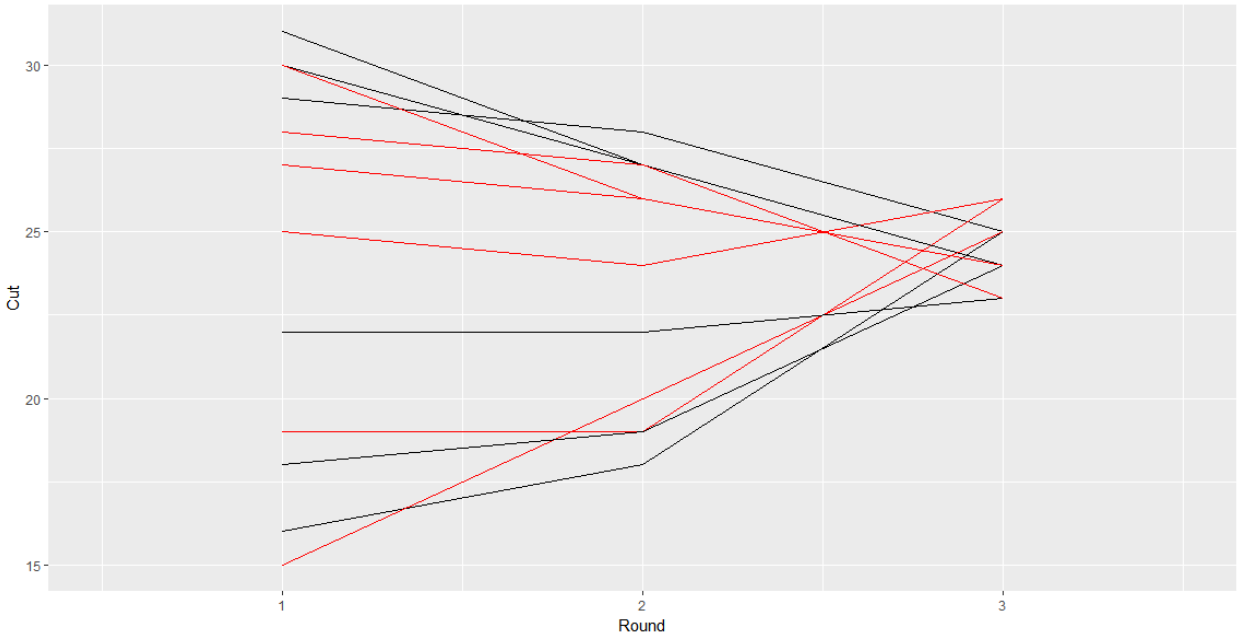


Figure 7.5 provides another example of a visualization reflecting the variability in the panelist's cut scores across rounds.

Figure 5.5. Across Round Trends in Panelist PL 2 Cut Scores



Note. Table 1 = —— Table 2 = ——

Table 5.2 presents an example of how the impact data (i.e., the percentage of students in the two performance levels) from each round will be presented. These percentages will be based on all students who took the Spring 2016 Grade 3 ELA assessment.

Table 5.2. Impact Data after Each Round

	PL 2		PL 3	
	Not Pass	Pass	Not Pass	Pass
Round 1				
Round 2				
Round 3				

***Note*. Table will be completed after the event.**

5.2. State Broad Review and Approval

The State Board meeting to review and approve the recommended cut score is scheduled for **TBD**. The final approved cut score and the resulting impact data are provided in Table 7.2 (Round 3) and Tables 1.2 and 1.3 in the Executive Summary (Section 1).

6. Reliability and Validity Evidence

6.1. Reliability Evidence

6.1.1. Intra-panelist Consistency

Intra-panelist reliability measures the degree to which panelists' ratings vary across rounds (Berk, 1996). Variability across rounds is expected to be relatively high if panelists are integrating information presented and making adjustments to their ratings. Consequently, intra-panelist reliability should yield a low coefficient when panelists make such adjustments. Higher coefficients are attained when many panelists do not modify their original decisions or make only minor adjustments. Intra-panelist consistency will be estimated by calculating intra-class correlation coefficients[15] (two-way random model, absolute agreement, for averages).

6.1.2. Inter-panelist Consistency

Inter-panelist reliability measures the degree to which ratings are consistent across panelists (Berk, 1996). Since panelists will be given the opportunity to adjust their cut scores based on various information sources including empirical data and group discussion, inter-panelist reliability estimates could be expected to increases across rounds. The inter-panelist reliabilities for each round will also be estimated by intra-class correlation coefficients[16] (two-way mixed model, absolute agreement, for averages). It is expected that the inter-panelist reliabilities will increasing across the rounds.

[15] The data frame was a matrix of panelists (rows) by the two rounds of cut scores (cols).
[16] The data frame was a matrix of OIB items (rows) by panelists (cols). Round 3 is not included as only a single cut score was given by panelists. The "mixed" design treats panelists as a "fixed" versus "random"

Another measure of inter-panelist agreement is the standard deviation (SD) and standard errors (SE) of the individual panelists' cut scores (Hambleton & Pitoniak, 2006). Table 6.1 presents an example of how these will be presented for each round within each cut score. Variability should decrease noticeably across rounds.

Table 6.1. Standard Deviations and Standard Errors

	PL 2			PL 3		
Index	**Round 1**	**Round 2**	**Round 3**	**Round 1**	**Round 2**	**Round 3**
SD						
SE						

***Note*. Table will be completed after the event.**

Visual evidence of rater variability may be seen in Figures #.#, #.#, and #.#.

6.2. Evaluation Survey

The evaluation survey (Appendix X) will be used to gather evidence about the validity of the standard's validation results. The form focused on the following areas to evaluate different aspects of the standards evaluation meeting:

1. Orientation Session
2. Group Discussions
3. Rating Activities
4. Influential Factors
5. Satisfaction with Questar Staff
6. Usefulness of Materials
7. Group Results for the PL 2 and PL 3 Cut Scores

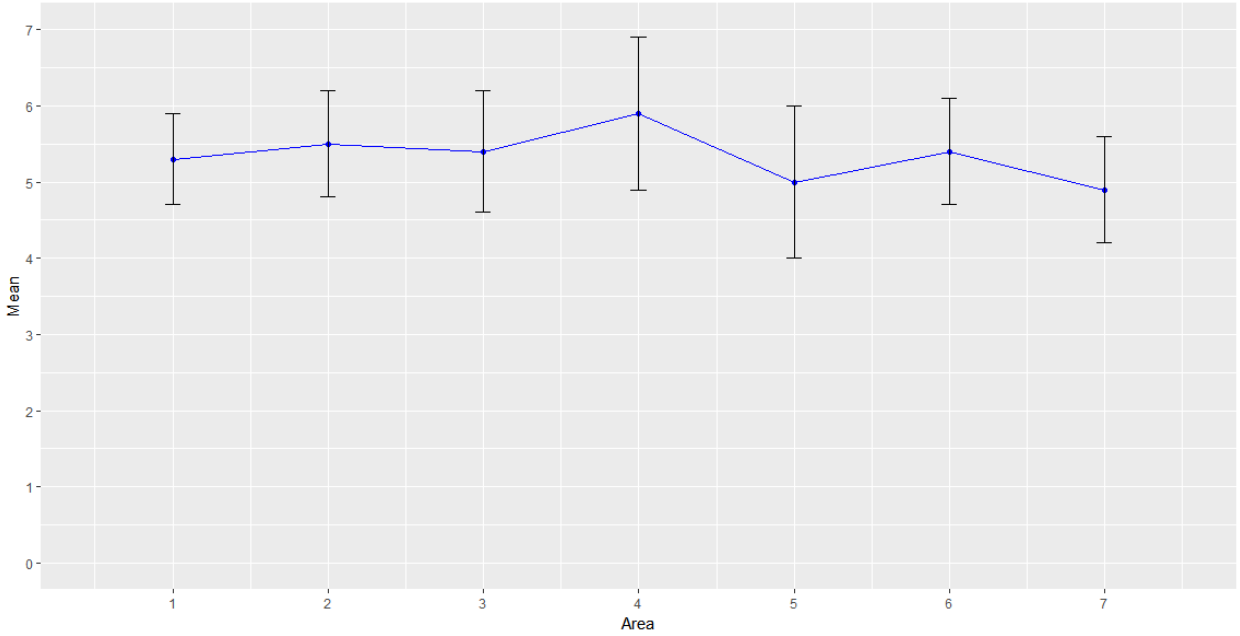
Each area will include several questions. Some response scales will be bidirectional and will use a six-option Likert-type scale with no middle option (e.g., "Strongly Disagree", "Disagree", "Slightly Disagree", "Slightly Agree", "Agree" and "Strongly Agree"). Other response scales will be unidirectional and include four response options. Three response option sets will be used for the four-option scales: (1) "Not Important", "Somewhat Important", "Important" and "Very Important"; (2) "Not Satisfied", "Partially Satisfied", "Satisfied" and "Very Satisfied"; and (3) "Not Useful", "Partially Useful", "Useful" and "Very Useful".

An appendix in the final report will document the full results of the evaluation form, which will be summarized, as shown below. For the statistical analysis, the response options will be coded as 1, 2, 3, 4, 5, and 6 and 1, 2, 3, and 4, respectively. For example, the mean responses for each evaluation strand for the four and six option scales will be presented, as demonstrated in Figures 8.1. With such plots one can easily locate areas that had the highest and lowest mean scores as well as which areas were

facet. Some might argue that "random" is more appropriate in this context; however, the pattern of results is not expected to change appreciably because of this.

the least and most variable. The responses to the individual questions within each area will also analyzed in a similar fashion.

Figure 6.1. Mean Responses to Each Evaluation Area with Six Options



The remaining questions in the evaluation survey will ask about the panelists' confidence in their final recommended Not Pass and Pass cut score. Specifically, these final statements will ask about the panelists' confidence in the reasonableness, appropriateness, and defensibility of the cut score. The goal will be for the majority of panelists' responses to fall in the 'Agree,' and 'Strongly Agree' categories. This will indicate that panelists were, as a whole, confident with the final recommended cut score.

Table 8.# Frequency Distribution of Reponses to the Reasonableness of the PL 2 *Not Pass and Pass* Cut Score

Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
0.00	#	#	#	#	#

***Note*. Table will be completed after the event.**

Table 8.# Frequency Distribution of Reponses to the Appropriateness of the PL 2 *Not Pass and Pass* Cut Score

Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
0.00	#	#	#	#	#

***Note*. Table will be completed after the event.**

Table 8.# Frequency Distribution of Reponses to the Defensibility of the PL 2 *Not Pass and Pass* Cut Score

Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
0.00	#	#	#	#	#

***Note*. Table will be completed after the event.**

6.3. Hambleton (2001) Validity Considerations

The purpose of the final report will be to provide comprehensive documentation of the validity evidence for this standards validation. Table 6.2 addresses validity considerations suggested by Hambleton (2001, p. 108 – 113). Preliminary responses based on current expectations are provided below and will be updated once the standard's validation is complete.

Table 6.2. Summary of Validity Evidence

Question	Response
Was consideration given to the groups who should be represented on the standards validation panel and the proportion of the panel that each group should represent?	Section X.X. provides detailed information about the panelists. When recruiting panelists, the MDE will consider several factors such as diversity in demographics, expertise in early elementary reading, and teaching experience.
Was the panel large enough and representative enough of the appropriate constituencies to be judged as suitable for setting performance standards on the educational assessment?	Table X.X provides panelists' self-reported demographical composition information to view the representativeness of each panel. The event planned for 12 panelists with a goal of having at least 10 panelists participating in each grade level test standards validation.
Were two panels used to check the generalizability of the performance standards across panels?	One panel will be used at this event. However, panelists will be spread across two tables.
Were sufficient resources allocated to carry out the study properly?	A one-day event will be employed. The panel will be led by an experienced psychometrician as the facilitator. Panelists will also have access to a content expert, and MDE staffers will be available as well. Most event documents are included in the appendices of this report so their quality can be inspected.
Was the performance standards validation method field tested in preparation for its use in the standards validation study? Was it revised accordingly?	No formal field test will be undertaken. However, the facilitator had conducted numerous prior standard setting and standard's validation events. The facilitator will be able to take advantage of their prior experiences with these two methods when planning this event. Procedures will also be reviewed by the state's TAC.
Was the standards validation method appropriate for the particular educational assessment? Was it described in detail?	Section X.X provides the rationale for using the OIYN method. This method will be approved by the MDE and vetted by their TAC. Details about the method and the procedures employed are provided throughout this report.

Question	Response
Were panelists explained the purposes of the educational assessment and the uses of the test scores at the beginning of the standards validation meeting?	The MDE and Questar will provide introductory information the morning of the event. Appendix X provides the orientation and training PowerPoint.
Were panelists exposed to the actual assessment? How it was scored?	Panelists will take all reading operational items before they make their ratings in order to become familiar with the items and test difficulties. Scoring information will be provided at the event.
Were the qualifications and other relevant demographic data about the panelists collected?	Table X.X presents the results from the panelist information form.
Were panelists administered the educational assessment, or at least a portion of it?	Panelists will take all reading operational items before they make their ratings.
Were panelists suitably trained on the method to set performance standards? For example, did the panelists complete a practice exercise?	Detailed training and practice activities will be used. Appendix X provides the orientation and training PowerPoint. Panelist will practice making yes/no decisions over a set of nonoperational items. The responses to the readiness form is expected to show that panelists understood the procedure and were ready to begin before starting Round 1.
Were descriptions of the performance categories clear to the extent that they would be used effectively by panelists in the standards validation process?	Appendix X provides the PLD that was described during the training. Content specialists and MDE staff were also available during the event to answer panelists' questions. The facilitator script illustrates that panelists were frequently reminded to consider the PLD in their decision-making. The evaluation survey asked for panelists feedback on PLD. Results showed all the panelists felt the PLD was useful to them during the standard's validation (see Section X.X).
If an iterative process was used for discussing and reconciling rating differences, was the feedback to panelists clear, understandable, and useful? Was the facilitator able to bring out appropriate discussion among the panelists without biasing the process?	Three rounds were used. Results from prior rounds were reviewed and discussed before panelists made the next round's ratings. The evaluation survey (Appendix X) asked for panelists to rate the effectiveness of the feedback information provided to them and from the group discussions. Results showed that all panelists felt that the Round 1 and 2 discussions (e.g., item difficulty and impact data) were helpful to them. See Section X.X for details. The facilitator script illustrates the types of open-ended questions (see Appendix X) panelists were asked in order to stimulate their discussions.
Was the process conducted efficiently? Were the item rating forms easy to use? Were documents such as examinee booklets, tasks, and items simply coded?	The responses to the evaluation survey were positive (see Section X.X), indicating that the process went efficiently. In addition, panelists had high confidence about the final recommended cut scores. All panelists agreed or strongly agreed that the item rating form was easy to use (see Section X.X).
Were documents such as examinee booklets, tasks, and items simply coded? If copies of examinee work was used, were they easily readable?	All documents were high-quality resolution prints, many in color, and reviewed by staff for legibility prior to packaging and distribution. Most event documents are included in the appendices of this report so their quality can be inspected.
Was the facilitator qualified?	**The panel had an experienced facilitator (e.g., 1) …..).**

Question	Response
Were panelists given the opportunity to "ground" their ratings with performance data? How was the data used?	The items were ordered by their item difficulties (Rasch values) and provided in the OIB. The item difficulties will be also graphically presented in a 'dot' plot (see Appendix X). The OIB will be used through three rounds, and the dot plot will be shown to panelists after Round 1. For more details, see the facilitator script in Appendix X.
Were panelists provided consequential data (or impact data) to use in their deliberations? How did they use the information? Were the panelists instructed on how to use the information?	Impact data was provided after all three rounds. Instructions on use was covered in training and explained by the facilitator (see the facilitator script in Appendix X).
Was the approach for arriving at final performance standards clearly described and appropriate?	Information provided to the panelists are covered in the training PowerPoint and the facilitator script.
Was an evaluation of the process carried out by the panelists?	An evaluation survey was completed by panelists. Section X.X summarizes the responses to each area of the process.
Was evidence compiled to support the validity of the performance standards?	
Was the full standards validation process documented?	This report documents all processes employed at the event. Appendices include materials such as the agenda, orientation and training PowerPoint, item rating form, evaluation survey, and the facilitator script.
Were effective steps taken to communicate the performance standards?	The State Board of Education approved the PLD (see Appendix X). These are provided on the state website. General descriptions are printed on score reports.

***Note*. Table will be completed after the event.**

6.4. Procedural Validity Evidence

The following section will be completed after the event.

Explicitness	The standard's validation plan and procedures will be vetted by the state's National TAC at two separate meetings held in advance of the event
Practicality	See the **Evaluation Form** results presented in (**TBD**)
Implementation	See the **Evaluation Form** results presented in (**TBD**).
Feedback	**Evaluation Form** results are presented in (**TBD**)
Documentation	This Technical Document

6.5. Internal Validity Evidence

Consistency within method	NA (only one method employed)
Intra-panelist consistency	See **TBD**. Figures **TBD** help visualize the variability.
Inter-panelist consistency	See **TBD**. Visual supplements include Figures **TBD**. As discussed elsewhere in this report, consistency generally improved across rounds for each performance level.
Decision consistency	**TBD**
Other measures	NA.

6.6. External Validity Evidence

External Validity Evidence

Comparison to other methods	NA—other standard's validation methods will be not implemented
Comparisons to other sources	Currently NA –
Reasonableness of cuts	The **Evaluation Form** asked panelists about the reasonableness of the final recommended cuts. The results showed that almost all the panelists agreed or strongly agreed about their reasonableness. See **Tables** **TBD**. Subgroup performance level percentages **TBD.**

7. References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*, *2nd edition* (pp. 508-600), Washington, DC: American Council on Education.

Cleveland, W.S. and McGill, R. (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79:531-554.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. Brennan (Ed.) *Educational Measurement, 4th Edition*, (pp. 433-470). Westport, CT: Praeger Publishers.

Impara, J. C. & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 12*, 355-368.

Mitzel, H. C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.). *Setting Performance Standards: Concepts, methods, and perspectives*, (pp. 249-282) Mahwah, NJ: Lawrence Erlbaum.

Smith, R. W., & Davis, S. L., (2009). *Combining the best of both worlds: The ordered item booklet Angoff*. Paper presented at the annual conference of the National Council on Measurement in Education, San Diego, CA Month, 2009.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, CA: National Academy of Education.

Skaggs, G., & Tessema, A. (2001, Month). *Item disordinality with the Bookmark standard setting procedure.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Appendix X: Agenda

Appendix X: Panelist Information

Note: **Will be completed after the event. Will be based on information from the Evaluation Form.**

Appendix X: Performance Level Descriptor (PLD)

Appendix X: Borderline Student Worksheet

Appendix X: Item Rating Form

Appendix X: Orientation PowerPoint Presentation

Appendix X: Facilitator Script

Appendix X: Evaluation Survey

Appendix X: Evaluation Survey Results

Note: **Will be completed after the event.**

Appendix X: Readiness Form