

Maine Through Year Assessment Spring 2023 Technical Report



© 2023 Maine Department of Education. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

Table of Contents

Section 1: Overview of the Maine Through Year Assessment	1
1.1. Intended Purposes and Uses of Test Results	1
1.2. Required Assessment and Policies for Including All Students	2
1.3. Meaningful Consultation.....	2
1.3.1. Schedule of Major Events.....	2
Section 2: Test Design and Development	4
2.1. Test Design & Development.....	4
2.1.1 Achievement Level Descriptors	5
2.2. Test Plan.....	6
2.2.1. Maine Revised Blueprint Explanation	7
2.2.2. Math Summative Blueprint Considerations	7
2.2.3. Reading Summative Blueprint Considerations.....	9
2.3. Item Development and Guidelines	13
2.3.1. Universal Design	16
2.4. Content and Bias Review.....	17
2.5. Data Review	18
Section 3: Administration and Security	21
3.1. Administration	21
3.2. Spring 2023 Administration	21
3.2.1. Student Population.....	21
3.2.2. Item Pool Characteristics.....	24
3.3. Constraint-Based Adaptive Test Engine.....	24
3.3.1. Engine Evaluation	26
3.3.2. Blueprint Constraint Accuracy	26
3.3.3. Score Precision	28
3.3.4. Item-Exposure Rates.....	30
3.3.5. Item Sequence	33
3.4. Paper Form Administration.....	33
3.5. Spring 2023 Fixed Form Blueprints	33
3.5.1. Receive and Take Inventory of School Materials	36
3.5.2. Score Transcription	36
3.6. Assessment Security.....	37
3.6.1. Assessment Ethics and Appropriate Practice	37
3.6.2. Online Security.....	37
3.6.3. Student Assessment Security.....	38
3.6.4. Returning or Destroying Secure Materials	38
3.7. Systems for Protecting Data Integrity and Privacy.....	38
Section 4: Item Statistics, Calibration, and Scaling.....	40
4.1. Classical Item Statistics	40
4.1.1. Expected P Value.....	40
4.1.2. Item Discrimination (Item-Total Correlation)	43
4.2. IRT Calibration.....	46
4.3. IRT Model Assumptions.....	47

4.3.1. Local Independence	47
4.3.2. Model Fit	47
4.3.3. Unidimensionality	48
4.4. Scaling	48
Section 5: Technical Quality-Validity	50
5.1. Validity Evidence Framework	50
5.2. Purposes and Evidence	51
5.2.1. Test Purpose 1	51
5.2.2. Test Purpose 2	52
5.2.3. Test Purpose 3	53
5.2.4. Test Purpose 4	54
5.3. Interpretive Argument Claims	54
5.4. Validity Argument	55
Section 6: Technical Quality-Other	57
6.1. Reliability	57
6.1.1. Marginal Reliability for Adaptive Tests	57
6.1.2. Reliability for HS Fixed Forms	58
6.1.3. Classification Accuracy	59
6.2. Fairness and Accessibility	63
6.2.1. Logistic Regression (LR) DIF Method	63
6.2.2. DIF Results	65
6.3. Full Achievement Continuum	69
6.4. Scoring	69
6.4.1. Construct Maine Scale	69
6.4.2. Machine-Scored Items	70
6.4.3. Attemptedness Rule and Not-Tested Codes	70
6.5. Multiple Assessment Forms	70
6.6. Multiple Versions of an Assessment	71
6.7. Technical Analysis and Ongoing Maintenance	71
Section 7: Inclusion of All Students	72
7.1. Testing Population	72
7.2. Procedures for Including Students Who Utilize Accessibility Features	72
7.3. Procedures for Including Multilingual Learners	72
7.4. Accommodations	73
7.5. Monitoring Test Administration for Special Populations	74
7.5.1. Monitoring in Acacia	74
7.5.2. Maine DOE Site Visits	75
Section 8: Achievement Standards and Reporting	80
8.1. State Adoption of Achievement Standards	80
8.2. Achievement Standard Setting	81
8.3. Reporting	84
8.3.1. Achievement Level Descriptors	84
8.3.2. Setting the Cut Scores	84
8.3.3. Reports	85

Section 9: References.....88

List of Tables

Table 1.1. Schedule of Major Events for the Spring 2023 Administration	2
Table 2.1. Number of Items and Points Per Test.....	7
Table 2.2. Math Blueprint Percentages, Grades 3–5.....	9
Table 2.3. Math Blueprint Percentages, Grades 6–8 & 10.....	9
Table 2.4. Reading Blueprint Percentages, Grades 3–8 & 10	10
Table 2.5. Reading Lexile Ranges, Grades 3–8 & 10.....	10
Table 2.6. Reading Word Count Ranges, Grades 3–8 & 10	10
Table 2.7. Reading Blueprint Percentages, Grades 3–8 & 10	13
Table 2.8. Math and Reading Online Item Types	15
Table 2.9. Item Type Percentages by Grade—Reading	15
Table 2.10. Item Type Percentages by Grade—Mathematics	16
Table 2.11. 2023 Content and Bias Review Results.....	18
Table 2.12. Data Review Flagging Criteria—Multiple-Choice and Non-Multiple-Choice Items...	19
Table 2.13. Data Review Results (forthcoming)	20
Table 3.1. Demographic Information—Reading.....	22
Table 3.2. Demographic Information—Mathematics.....	22
Table 3.3. Ability Distribution.....	24
Table 3.4. Number of Items by Content Category (Spring 2023 Summative Item Pool).....	24
Table 3.5. Blueprint Constraint Accuracy by Reporting Category	27
Table 3.6. CSEMs at the Cut Scores.....	29
Table 3.7. CSEMs by Score Decile	29
Table 3.8. Operational Item Exposure Rates.....	31
Table 3.9. Field Test Item Exposure Rates	32
Table 3.10. Paper Form Summative Item Totals by Content and Grade.....	33
Table 3.11. Reading Item Counts by Instructional Area, Grades 3–8	34
Table 3.12. Reading Item Counts by Instructional Area, Grade 10	35
Table 3.13. Mathematics Item Counts by Instructional Area, Grades 3–5	35
Table 3.14. Mathematics Item Counts by Instructional Area, Grades 6–8	35
Table 3.15. Mathematics Item Counts by Instructional Area, Grade 10	36
Table 4.1. Summary of <i>P</i> Values—Operational Items	41
Table 4.2. Summary of <i>P</i> Values—Field Test Items	42
Table 4.3. Summary of Item-Total Correlations—Operational Items.....	44
Table 4.4. Summary of Item-Total Correlations—Field Test Items	45
Table 4.5. Summary of IRT Item Statistics—Operational Items.....	46
Table 4.6. Summary of Mean-Square Infit and Outfit Statistics	48
Table 4.7. Maine Grade-Level Scale Properties.....	49
Table 5.1. Sources of Validity Evidence for Each Test Purpose	51
Table 5.2. Interpretive Argument Claims, Evidence to Support the Essential Validity Elements	54
Table 6.1. Reliability Statistics—Reading	58
Table 6.2. Reliability Statistics—Mathematics	58
Table 6.3. Cronbach’s Alpha (Internal Consistency) for Fixed Forms	59
Table 6.4. Classification Accuracy by Achievement Level—Reading	60
Table 6.5. Classification Accuracy by Achievement Level—Mathematics.....	61
Table 6.6. LR DIF Categories.....	65

Table 6.7. DIF Analysis Results—Operational Items.....	65
Table 6.8. DIF Analysis Results—Field Test Items.....	67
Table 6.9. Available Not-Tested Codes.....	70
Table 7.1. Number of Students Who Used TTS	74
Table 8.1. MTYA Achievement Level Descriptors	80
Table 8.2. Final Approved Cut Scores—Reading.....	83
Table 8.3. Impact Data Associated with Cut Scores—Reading	83
Table 8.4. Final Approved Cut Scores—Mathematics	83
Table 8.5. Impact Data Associated with Cut Scores—Mathematics	83
Table 8.6. Report Levels	85

List of Figures

Figure 2.1. Test Development Process	5
Figure 2.2. Maine Blueprint Percentages—Math, Grades 3–8 & 10	8
Figure 2.3. Passage Quality Checklist.....	11
Figure 2.4. Internal Item-Development Process Overview.....	14
Figure 3.1. Adaptive Engine Overview	25
Figure 3.2. Student-Specific Plan Approach.....	26
Figure 7.1. Monitoring Testing Status in Acacia	75
Figure 7.2. 2023 Maine Through Year Assessment Observation Form	76
Figure 8.1. Maine Through Year Assessment Embedded Standard Setting Iterative Processes	81
Figure 8.2. Individual Student Report—Page 1	86
Figure 8.3. Individual Student Report—Page 2	87

Section 1: Overview of the Maine Through Year Assessment

This technical report documents the processes and procedures implemented to support the Maine Through Year Assessment program managed by NWEA under the supervision of the Maine Department of Education (DOE). The Through Year Assessment includes assessments in reading and mathematics for grades 3 through 8 and the second year of high school (grade 10). This technical report shows how the processes, methods applied, and results relate to the issues of validity and reliability and to the *Standards for Educational and Psychological Testing* (AERA et al., 2014). The complete technical report will be made available to the public by the Maine Department of Education at https://www.maine.gov/doe/Testing_Accountability/MECAS/NWEA no later than February 15, 2024.

The Maine Through Year Assessment is mostly an online adaptive test. For students with a need documented in an Individualized Education Plan (IEP) or 504 Plan, the test also offers three accommodated paper forms: paper/pencil standard print forms, large print forms, and braille forms. There are three administrations of the Through Year Assessment: fall, winter, and spring. The fall and winter administrations are diagnostic tests that are used to measure and predict student growth. The spring administration is a combination of the state summative test and the diagnostic test, with the summative test making up the majority of the assessment. The state summative test is designed to fulfill peer review requirements and, for the purpose of this document, only the summative portion of the assessment will be discussed.

Spring 2023 was the first administration of the Maine Through Year Assessment. This report focuses on the processes and procedures related to the state summative test to comply with the peer review guidance. In Spring 2023, post-equating and standard setting were conducted on the state summative test to construct the Maine scale score and achievement levels. The design of the state summative test, psychometric analyses, test validity, reliability, and standard setting are described in various sections in this report.

1.1. Intended Purposes and Uses of Test Results

The Maine Through Year Assessment has four primary purposes:

1. To report individual student achievement relative to the state-adopted content standards in reading and mathematics
2. To provide information to the public about school performance through the state's Every Student Succeeds Act (ESSA) reporting system, the ESSA Data Dashboard
3. To support school identification within the state's ESSA compliant system of school identification and support
4. To provide a source of information for ongoing local program evaluation

The Maine Through Year Assessment is designed to measure Maine's accountability standards, the Common Core State Standards (CCSS) in math and reading. Student results are reported according to academic achievement descriptors utilizing cut scores established in embedded standard setting for each of four achievement levels: *Well-Below State Expectations*, *Below State Expectations*, *At State Expectations*, *Above State Expectations*.

1.2. Required Assessment and Policies for Including All Students

Students in grades 3-8 and second year of high school participate in the Maine Through Year Assessment. Students with disabilities and multilingual learners may participate in the Maine Through Year Assessment with accommodations.

Exceptions to participation would occur in cases involving students with the most significant cognitive disabilities who have been found eligible for alternate assessments via the IEP Team Process. Only about 1% of all publicly funded Maine students in grades eligible for assessment participate in an alternate assessment; the rest of the student population (approximately 99%) participate in the Maine Through Year Assessment.

1.3. Meaningful Consultation

1.3.1. Schedule of Major Events

Table 1.1 presents the major events that occurred for the 2023 Maine Through Year Assessment.

Table 1.1. Schedule of Major Events for the Spring 2023 Administration

Event	Date(s)
ALD Workshop	September 12–13, 2022
Content and Bias Review	December 1–2, 2022
Alignment Study/Embedded Standard Setting	Reading: July 18–20, 2023 Math: July 25–27, 2023
Test Administration Training ^a	March 16 and 21, 2023
Operational Test Window	May 1–26, 2023
Data Review	October –November 2023
Technical Advisory Committee Meeting	October 12, 2022 January 25 and 30, 2023 (half days) August 18, 2023

^a Test Administration Training slides are included in Appendix A.

This list provides more details about the events presented in the table.

- Achievement Level Descriptor (ALD) Workshop: a workshop with Maine educators to review and refine language in the Achievement Level Descriptors, ensuring cohesion within and across grade levels
- Content and Bias Review: a meeting with Maine educators to review all items authored for the program by NWEA
- Alignment Study/Embedded Standard Setting: an educator review of selected Maine items to their standards by a third-party vendor, followed by standard setting by another third-party vendor to establish cut scores at each achievement level
- Test Administration Training: training to prepare District Assessment Coordinators, School Assessment Coordinators, and proctors. Topics covered include Through Year Assessment Overview, Assessment Management in Acacia, Accessibility, Not-Tested Codes, Preparing and Monitoring the Assessment, Regional and Out-of-State Programs,

Proctor/Student Experience, Operational Reports, Data and Reporting, Preparation, Resources and Tips, and Communication and Partner Support.

- Operational Test Window: the time period that Maine students take the summative assessments
- Data Review: a review/analysis of field test items that were flagged for item performance. NWEA shares/discusses with MDE the results of this review, and decisions are made regarding the next steps for the flagged items
- Technical Advisory Committee (TAC) Meeting: a meeting with selected and designated assessment experts to review, discuss, and advise Maine's assessment program. Additional TAC member information and meeting topics can be found in Appendix I.

Below is a list of topics from the TAC meetings leading up to the Spring 2023 Through Year Assessment administration:

- October 12, 2022
 - Program Overview
 - Test Blueprints
 - Test Design
 - Achievement Level Descriptor Workshop
 - Embedded Standard Setting
- January 25, 2023
 - Test Design
 - Equating Plan
- January 30, 2023
 - Equating Plan (continued from Jan 25 meeting)
 - Score Report Mockups
 - Comparability Study

Section 2: Test Design and Development

This section describes the test design and development processes for the Spring 2023 Maine Through Year Assessment.

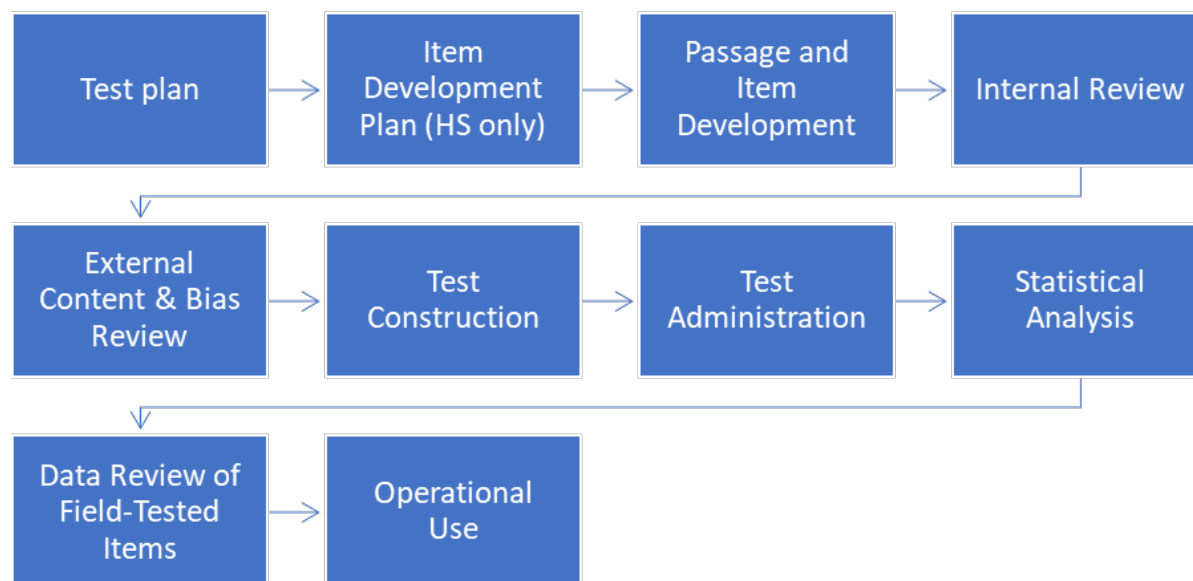
2.1. Test Design & Development

Maine administered computer adaptive assessments in reading and mathematics for grades 3–8 and a fixed form in reading and mathematics for the second year of high school (since there was no item pool for an adaptive model). Items in the grades 3–8 adaptive assessments were licensed from NWEA. For the grade 10 (HS) assessments, passages were identified or developed, and items were developed for the assessments.

Items were written internally by NWEA content specialists or external professional item writers. Items may align to part of one standard or to the entire standard. NWEA’s content specialists identify the cognitive demand of each item. The rigor of each item comes from the knowledge and skills required to answer the item correctly, which is dictated by the standard. Test items may be written to Webb’s Depth of Knowledge levels—1, 2, or 3. A particular content standard that has been unpacked (broken down) can have a pool of items developed to the unpacked parts of the standard that align to more than one DOK level. Developers of the summative tests, however, also aim to develop items that assess a standard/Achievement Level Descriptor (ALD) as a whole to determine a student’s complete mastery of that standard/ALD. Additionally, content developers ensure that summative items meet the ALDs (what a student should be able to do at a particular grade level regarding on-grade-level content). Once developed, items and passages go through multiple (and thorough) reviews for content, bias and sensitivity, permissions, editorial fidelity, and item functionality.

All newly developed grade 10 items were submitted to a virtual Content and Bias Review meeting with Maine educators held in December 2022. Educators reviewed all items that had been developed for the high school assessments and designated each item as “accept,” “accept with revisions,” or “reject.” Then educators discussed their reviews and came to a consensus for each item; final decisions were recorded by the meeting facilitator. Following the meeting, NWEA assessment specialists applied any necessary revisions to the items. Items were then considered for the field test pool. Please note that these newly developed items also served in an operational capacity, since there was no existing high school item pool. Figure 2.1 outlines the general steps taken to develop the passages and items for use on the high school assessments.

Figure 2.1. Test Development Process



2.1.1 Achievement Level Descriptors

Range Achievement Level Descriptors (ALDs) show a progression of skills within a standard over multiple achievement levels. Range ALDs describe what a student should likely be able to do at a particular achievement level regarding on-grade content. For each assessed standard, the ALDs show the range of on-grade content from easiest, or least cognitively challenging, to most difficult, or most cognitively challenging.

The intent is that the ALDs, when viewed as a whole, provide a wide range of knowledge, skills, and abilities students can demonstrate over the course of the year while also considering the work from the previous grade and the upcoming work in the next grade. Some content may appear in multiple places in the standards, but the ALDs are written to minimize overlap between grades. For example, CC math standards 3.NBT.1 and 4.NBT.3 both assess rounding whole numbers. The ALDs for these standards use grade-level content limits to ensure that an item assessing rounding will only align to one grade.

Range ALDs allow students at various levels to demonstrate their knowledge and skills. Range ALDs allow for more adaptivity during a test event based on each student's individual performance. Range ALDs help describe a student's current level of understanding, which allows stakeholders the achievement to pinpoint areas of strength and areas of growth. Range ALDs are also used to guide NWEA content specialists in writing items for assessments.

NWEA content specialists wrote the initial draft of the Maine Range ALDs and then held a workshop with Maine educators to review and revise the ALDs. Maine educators were asked to review these NWEA ALDs in relation to the Common Core State Standards used in Maine. Each participant reviewed Range ALDs for grades 3–8 and second year of high school

(grade 10) in either reading or mathematics. The review’s purpose was to give Maine educators an opportunity to study the ALDs and share their feedback with NWEA content specialists.

The number of committee members for each content area was limited to two–three educators. For this reason, educators with expertise in all grade levels were recruited to participate. The state identified approximately 140 curriculum coordinators. The DOE emailed these educators a link to a survey generated by NWEA to indicate their interest and availability. Seven educators with positions as district administrators or curriculum specialists responded and were invited to participate. Of these seven educators, two declined and one did not complete the prework or attend the workshop. The four remaining participants represented three different regions of the state, including Southern Maine, Southern-Central Maine, and Down East Maine, and one educator represented a virtual academy. All participants had experience working in schools with a high number of economically disadvantaged students. Some participants had experience working with special education, English language learner, and gifted and talented students.

Maine educators were asked to complete prework for the ALD workshop. They were provided with a guide that defined Range ALDs, explained how they are organized, and described how they are used. The guide also outlined the review process and listed three statements to consider when evaluating the ALD progression for each standard. Each educator was given a version of the ALDs with two columns for feedback. The first column was used to indicate if they approved the ALD or would like to discuss the ALD at the workshop; the second column was used for comments.

NWEA content specialists compiled the feedback into one document and used it to determine which standards to discuss at the workshop. The NWEA content specialists also discussed the feedback with their content team before the workshop and suggested revisions to share with the educators. The workshop was held on the evenings of September 12–13, 2022. All standards marked by educators for discussion were addressed at the workshop. Four NWEA content specialists attended the workshop. Each content area had a content specialist that facilitated and another to help encourage discussion and record notes.

Both the reading and mathematics ALDs had progressions updated based on feedback from the Maine educators. These updates included reassigning ALD statements to another level within the progression, removing ALD statements, revising ALD statements, and crafting new ALD statements.

2.2. Test Plan

As part of test planning, decisions about how many operational items, how many field test items, and what standards would be assessed needed to be considered and finalized. Table 2.1 details the total item counts and the number of raw score points for the Spring 2023 summative test. Items administered included both operational and embedded field test items. High school students were administered a 30-item fixed form for the summative test. Items used for scoring in the high school grade were operational field test items.

Table 2.1. Number of Items and Points Per Test

Grade	Operational		Field Test	
	#Items	#Points	#Items	#Points
Reading				
3	27	30–31	7	7–11
4	27	30–31	7	7–13
5	27	30–31	7	7–14
6	27	30–31	7	7–14
7	27	30–31	7	7–14
8	27	30–31	7	7–14
HS	30	41	5	8
Mathematics				
3	27	30–31	7	7–10
4	27	30–31	7	7–10
5	27	30–31	7	7–10
6	27	30–31	7	7–10
7	27	30–31	7	7–11
8	27	30–31	7	7–10
HS	30	34	5	7

Note. HS items were operational field test items.

2.2.1. Maine Revised Blueprint Explanation

The NWEA Through Year summative blueprints outline the number of operational items that should be included on each test and the standards they are aligned to. For Maine, these summative blueprints include content categories consistent with the MAP Growth assessment in reading and mathematics and are aligned to the Common Core State Standards (CCSS), as required by the state. The content categories have been revised to provide consistency across all test administrations and with MAP Growth CCSS assessments. Each content category is weighted in the summative assessments based on the content category and the accountability needs for the state based on the standards assessed.

The blueprints were developed based on the priorities in the CCSS standards for both reading and mathematics. Although MAP Growth content categories are weighted equally due to diagnostic adaptability, the summative assessment content categories reflect the prioritization recommended for the Common Core State Standards. The percentage and percentage ranges reflect the standards in each content category in relation to the overall length of the test. All content categories are approximate and dependent on the total number of items per test and will require rounding if the item total by content category does not result in a whole number. Additionally, since the grades 3–8 tests are adaptive, the blueprint is programmed into the constraint-based engine (CBE) so that the requirements are met.

2.2.2. Math Summative Blueprint Considerations

The mathematics blueprints reflect the instructional emphasis of the content at each grade. For example, “Geometry” receives more instructional time as the grade levels progress and as the weight percentage increases from about 14% in grade 3 to about 28% in grade 10 (i.e., the second year of high school), as shown in Figure 2.2.

Figure 2.2. Maine Blueprint Percentages—Math, Grades 3–8 & 10

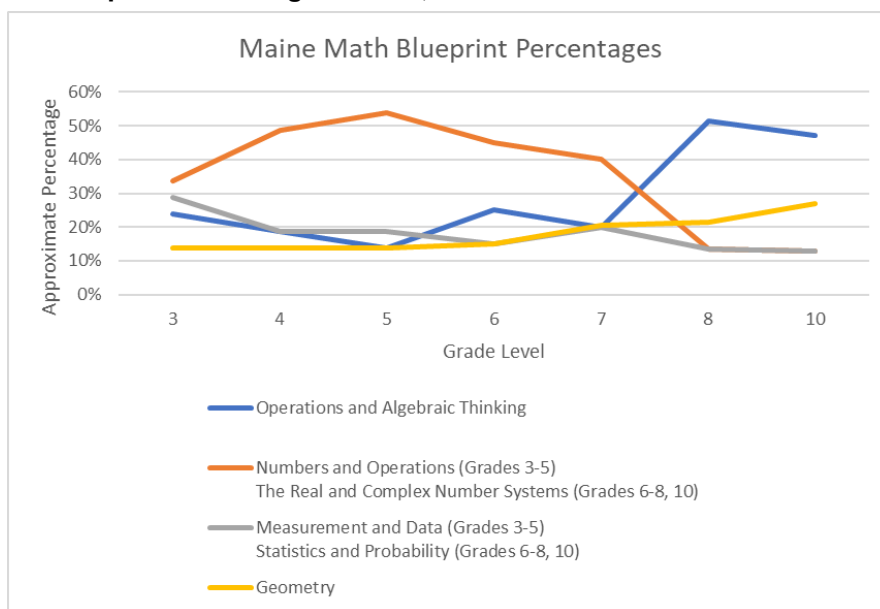


Figure 2.2 also shows that students’ skills in the “Numbers and Operations” content category progress as they work with whole numbers less than 1,000 and fractions with a limited set of denominators in grade 3 to decimals and a larger set of fractions in grade 5. After students grasp these skills, the significance of the content category (called “The Real and Complex Number Systems” starting in grade 6) gradually lessens as students work with the set of rational numbers in grade 6 to the set of irrational numbers in high school.

Conversely, students’ skills in the “Operations and Algebraic Thinking” content category steadily increase as students solve simple two-step problems in context in grade 3 to working with linear and quadratic functions in high school.

For the third content category (i.e., “Measurement and Data” in grades 3–5 and “Statistics and Probability” in grades 6–8 and 10), the percentage remains relatively constant and ranges from 10 to 30 percent. Students’ skills gradually progress as they work with picture graphs in grade 3 to scatter plots in high school.

For the “Geometry” content category, the percentage gradually increases from 15 percent in grade 3 as students work with area and perimeter to nearly 30 percent in high school as students work with more complex figures and geometric proofs.

Three grade 7 content categories are each assessed at approximately 20% in the blueprint because it is the point at which the “Operations and Algebraic Thinking” content category and the “Geometry” content category continue to increase, and the third content category (i.e., “Measurement and Data” in grades 3–5 and “Statistics and Probability” in grades 6–8 and 10) remains relatively constant near 20%.

Table 2.2 and Table 2.3 show the approximate percentages for the content categories for each math grade. Appendix B provides more detailed information about standard coverage, and Appendix G provides additional information about the blueprints.

Table 2.2. Math Blueprint Percentages, Grades 3–5

Content Category	Grade 3	Grade 4	Grade 5
Operations and Algebraic Thinking	23–25%	18–20%	13–15%
Numbers and Operations	33–35%	48–50%	53–55%
Measurement and Data	28–30%	20%	20%
Geometry	13–15%	13–15%	13–15%

Table 2.3. Math Blueprint Percentages, Grades 6–8 & 10

Content Category	Grade 6	Grade 7	Grade 8	Grade 10
Operations and Algebraic Thinking	25%	20%	48–53%	46–50%
The Real and Complex Number Systems	45%	40%	13–15%	13–15%
Geometry	15%	20%	21–23%	26–30%
Statistics and Probability	15%	20%	13–15%	13–15%

2.2.3. Reading Summative Blueprint Considerations

When creating the English language arts blueprints for Maine, the focus was on the weight and breadth of the reading standards designed to assess literary and informational texts and vocabulary skills (writing, language knowledge and conventions, and speaking/listening standards will not be assessed). Similar to the Priority Instructional Content guidance from Student Achievement Partners ([Achieve the Core](#)), the blueprint represents the belief that not all content standards are “emphasized equally” in the classroom and on assessments. In order to keep the text at the center and use text-based questions, these test items highlight close reading skills, text analysis, textual evidence, and academic vocabulary.

In addition to measuring skills and knowledge, a factor specific to English language arts assessments is the content of the passages used for test questions. There are two factors that directly affect the content of passages: (1) a balance of reading text content between literary and informational texts and (2) a range of text complexity. According to the Common Core State Standards, students are expected to demonstrate understanding of increasingly complex texts as a result of grade-level and discipline-specific content expectations.

Reading text content is classified as either literary or informational. The balance of percentages shifts from more literary content to more informational content as the grade-level increases. These percentages originated for grade bands with the Common Core State Standards and have been extrapolated to be grade-specific for Maine.

Table 2.4 shows the ratio of literary to informational text by grade.

Table 2.4. Reading Blueprint Percentages, Grades 3–8 & 10

Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Literary 55–60%	Literary 55–60%	Literary 50%	Literary 45–50%	Literary 40–45%	Literary 40–45%	Literary 40–45%
Informational 40–45%	Informational 40–45%	Informational 50%	Informational 45–50%	Informational 55–60%	Informational 55–60%	Informational 55–60%

Text complexity is the level of reading difficulty in order for students to understand what is read. A text complexity measurement is the process of evaluating a text for quantitative data, qualitative data, and the considerations for the reader and task. For items on the reading blueprint, students will encounter a range of text complexity within a grade level. Within a grade, text complexities should vary to include minimally complex, moderately complex, and highly complex.

Quantitative data includes concrete measures such as word length or frequency, sentence length, text cohesion, and vocabulary. These are communicated through readability measures to include Lexile, Word Count, and Flesch-Kincaid. Quantitative measures are a guide; exceptions can be made if the qualitative measures and/or grade-level alignments are appropriate. Table 2.5 shows acceptable Lexile ranges for each grade.

Table 2.5. Reading Lexile Ranges, Grades 3–8 & 10

Grade(s)	Lexile Range
3	450L–790L
4–5	745L–980L
6–8	925L–1155L
10	960L–1305L

Note. These Lexile bands reflect the adaptive nature of the assessments and the need to include a slightly larger range of readabilities than outlined in the [CCSS](#).

Table 2.6 provides acceptable word count ranges for each grade. For paired passages, each individual passage should fall within the word count range.

Table 2.6. Reading Word Count Ranges, Grades 3–8 & 10

Grade	Word Count Range
3	200–700
4	200–900
5	300–1000
6	400–1100
7	400–1100
8	400–1200
10	600–1400

Qualitative data includes the following dimensions: meaning/purpose, structure, language, and knowledge demands. Additionally, considerations regarding the reader and their interaction with a passage and the items they will answer for each passage help acknowledge students' role in the assessment. NWEA conducts a review of each passage using a Passage Quality Checklist (Figure 2.3) that determines the complexity and suitability for assessment.

Figure 2.3. Passage Quality Checklist

Passage Quality Checklist				
Title:	Author:	Grade Level or Band:		
Lexile:	FK:	Word Count:		
Selection Criteria		Comments		
1. Work worthy of study: a. Accurate content b. Lends itself to a close reading and analysis c. Provides ample opportunity for examining an author’s craft: i. Clear and effective structure ii. Development of arguments, ideas, characters, plot, setting (etc.) are detailed and thorough rather than superficial iii. Relevant evidence, reasoning, and concrete details iv. Rich, varied language (style, syntax, diction, rhetorical devices, domain-specific terms)		<input type="checkbox"/> Yes	<input type="checkbox"/> No	<input type="checkbox"/> Maybe
2. Free of bias and sensitivity concerns: a. Does not provoke an undue emotional response outside of highly individualized experiences b. Represents groups fairly, accurately, respectfully, and without stereotype c. Distinguishes traditional behaviors/values from stereotypes d. Presents differences and varieties without moral judgment e. Does not overgeneralize f. Characters are not depicted as victims of/dependent on dominant culture for help/success		<input type="checkbox"/> Yes	<input type="checkbox"/> No	<input type="checkbox"/> Maybe
3. Engaging and appropriate for target readers: a. Topics, issues, or arguments are likely to be of interest; OR b. Text is engaging		<input type="checkbox"/> Yes	<input type="checkbox"/> No	<input type="checkbox"/> Maybe
4. Ideal for assessment: a. Presents multiple opportunities for reading-related questions b. Appropriate for grade level given both text complexity and grade-specific standards c. Aligned to Georgia standards		<input type="checkbox"/> Yes	<input type="checkbox"/> No	<input type="checkbox"/> Maybe

5. Complex text that feels complete:	<input type="checkbox"/> Yes	<input type="checkbox"/> No	<input type="checkbox"/> Maybe
a. Appropriate for grade level or grade band based on quantitative and qualitative measures			
b. Does not require more prior knowledge than would be appropriate at the given grade			
c. Has the sense of a beginning, middle, and end.			
d. Does not require an elaborate contextual introduction			
e. Falls within word count guidelines for grade level or band (with allowance for +/-10%)			

For more information about text complexity, see <https://achievethecore.org/page/2725/text-complexity>.

Content categories are aligned to the prioritized standards into the following categories: literary text, informational text, and vocabulary.

Table 2.7. shows the approximate percentages for the content categories for each grade.

Table 2.7. Reading Blueprint Percentages, Grades 3–8 & 10

Content Category	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Literary Text	45–50%	40–45%	35–40%	35–40%	30–35%	30–35%	30–35%
Informational Text	30–35%	35–40%	35–40%	40–45%	45–50%	45–50%	45–50%
Vocabulary	20–25%	20–25%	20–25%	20–25%	20–25%	20–25%	20–25%

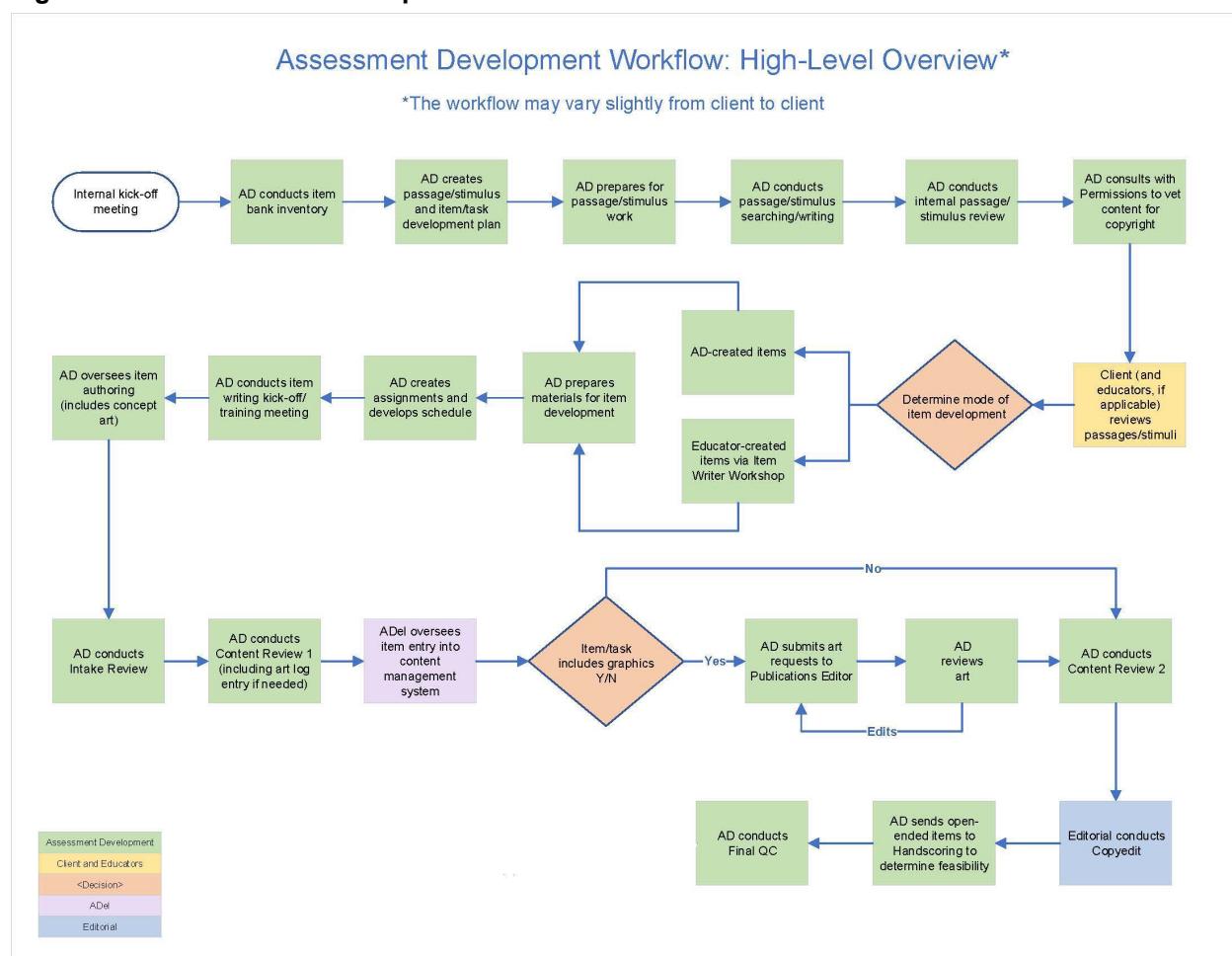
2.3. Item Development and Guidelines

Presenting students with solid test questions written in a consistent manner at an appropriate level that fairly assess the standards is critical. The key to achieving this result is using a team of seasoned writers to develop items and then following a clearly defined process that uniformly reviews and finalizes all items and passages/stimuli. NWEA follows consistent item-development and review processes that allow its team of experts (as well as educators) to revise, edit, and polish each item so that it is clearly understood and fair to all students, giving everyone an equal chance when answering each question.

NWEA follows a solid item-development process for all test items that assess standards (see Figure 2.4 for a high-level process overview). This process begins with an inventory (or plan) to identify the items that need to be developed and the specifications to which items will be written and ends with an item that will be field tested. Items are written by experienced writers who adhere to best practices for item development. Once items are written, they undergo extensive reviews to ensure they are appropriate for assessing the academic content standards. (For Maine, only grade 10 had newly development items.)

All newly written items undergo rounds of internal content and editorial review and revision. Items are reviewed for proper alignment and sound technical quality during internal and external reviews before being field tested. During this internal review, NWEA’s seasoned team of assessment specialists evaluate the items for alignment to the standards and item specifications, review the items’ rigor, and evaluate the items’ technical quality. Following the content reviews, NWEA’s editorial team copyedits the items. Next, items are reviewed by Maine educators during a Content and Bias Review meeting (external review). Finally, NWEA reviews and applies final revisions based on educator edits/suggestions. Once complete, items are added to the NWEA item pool and are ready for field testing. Item review checklists can be found in Appendix J.

Figure 2.4. Internal Item-Development Process Overview



For Maine, there was no targeted development for grades 3–8 for 2023, since NWEA had enough items for those grades in the pool. However, because there was no available grade 10 item pool, NWEA wrote items in-house for grade 10 reading and math. These items functioned as operational field test items in the grade 10 test. Item writers who were familiar with both industry best practices and NWEA’s standard item-development practices and processes were provided item specifications and an overview of the project (item specifications can be found in Appendix K). Writers followed Universal Design for Learning principles during item development.

Once the test plan was completed, a passage development/needs plan and an item-development plan were created. Some passages were written in-house, while others were located in the public domain. Passages were developed or selected in order to:

- offer appropriate content, length (emphasis on word counts), and text complexity
- provide engaging reading opportunities for students as they take the test
- include ample variation to appeal to a wide range of student audiences
- contain the characteristics required for the development of items that target a range of standards

Once reading passages were developed, selected, and/or approved, reading and math item development began based on needs outlined in the test blueprints. All content was reviewed during the process outlined above. After NWEA completed passage and item development, all items and passages were reviewed by Maine educators at a Content and Bias review (outlined in Section 2.4).

The Maine Through Year Assessment consists of several item types, as outlined in Table 2.8.

Table 2.8. Math and Reading Online Item Types

Item Type	Description
Multiple Choice	Students select one response from multiple options.
Multiselect	Students select two or more responses from multiple options. Some multiselect items are also two-point items for which students can earn partial credit.
Composite	Students interact with multiple interaction types included within a single item. Students may receive partial credit for composite items.
Gap Match	A type of drag-and-drop item in which students select one or more answer options from the item toolbox and populate a defined area, or “gap.” Some gap match items are also two-point items for which students can earn partial credit.
Graphic Gap Match	A type of drag-and-drop item in which students move one or more answer options from the toolbox and populate a defined area, or “gap,” that has been embedded within an image in the item response area. Some graphic gap match items are also two-point items for which students can earn partial credit.
Hot Text	Students select a response from within a piece of text or a table of information (e.g., word, section of a passage, number, symbol, or equation) that is highlighted in the selected text. Some hot text items are also two-point items for which students can earn partial credit.
Text Entry	Students input numeric answers using a keyboard.

Table 2.9 and Table 2.10 outline the percentages of item types by content area and grade level.

Table 2.9. Item Type Percentages by Grade—Reading

Grade	Item Type				
	Multiple Choice	Multiselect	Composite	Gap Match	Hot Text
2	84	16	0	0	0
3	83	7	4	4	1
4	86	6	4	3	0
5	85	8	2	3	2
6	87	7	3	2	1
7	77	11	5	4	2
8	84	6	4	4	2
10	87	10	3	1	0

Table 2.10. Item Type Percentages by Grade—Mathematics

Grade	Item Type						
	Multiple Choice	Multiselect	Composite	Gap Match	Graphic Gap Match	Hot Text	Text Entry
2	49	9	0	12	4	0	26
3	50	9	9	10	6	4	12
4	52	10	7	8	7	5	11
5	50	8	12	8	7	3	12
6	57	6	9	9	2	6	11
7	56	7	7	8	1	9	13
8	56	7	7	8	3	9	10
10	47	18	9	13	1	11	2

2.3.1. Universal Design

Ensuring that assessments are accessible to students with a variety of needs, including those with disabilities, is a critical part of item development. With a strong foundation in Universal Design for Learning (UDL) (Rose & Meyer, 2006), the assessments become engaging and accessible for all students. The NWEA content team ensures that each item is created with the principles of UDL in mind. These principles provide a framework for developing flexible items to support many kinds of learners and maximize options for assessments that provide multiple means of representation, action and expression, and engagement. Considerations NWEA takes into account when developing items include:

- Items are free of unnecessary linguistic complexity.
- Information presented in items is clear, concise, and relevant to the standard being assessed.
- Context and language are fair and familiar to students at their grade level and do not give advantages or disadvantages to subgroups.
- Items are free of stereotypes and potential disrespect regarding age, gender, race, ethnicity, language, religion, sexual orientation, social economic status, disability, or geographic region.
- Items do not challenge personal beliefs or values and avoid emotionally charged topics.
- Names and gender are avoided unless necessary. If names must be used, a variety of genders and ethnicities are represented.
- Graphics are intentional and not merely decorative.
- Graphics are not color dependent.
- MathML uses equation tags compatible with text-to-speech and screen readers.
- Art is tagged to be compatible with screen readers where possible.

Applying UDL principles to assessments helps reduce barriers and minimize irrelevant information from the items so the assessment can more appropriately capture what each student knows. It also ensures that there will be available items for the creation of accommodated forms, including large print and braille forms.

Items in the grades 3–8 item pool and items developed for the high school assessments were developed using the principles of UDL that are based on the notion that a good test design ensures optimal, standardized conditions to facilitate the reliability and validity of inferences regarding student achievement. The design of the assessments should be usable by all students to the greatest extent possible, including having tasks that are free of bias and construct-irrelevant content and are accompanied by clear and precise testing directions. An assessment should:

- measure what it intends to measure and reflect the intended content standard
- respect the diversity of the assessment population
- have a clear format for the text
- have clear pictures and graphics, including only essential illustrations
- have concise and readable text
- be amenable to accommodation
- minimize skills required beyond those being measured
- be accessible to all students (age, gender, ethnicity, disability, and socio-economic level)
- avoid content that might unfairly advantage or disadvantage any student subgroup

Once items were drafted, an assessment specialist reviewed the draft items for alignment to the target standard, item specifications, and style. The assessment specialist also verified that:

- The rigor of each item comes from the knowledge and skills required to answer the item correctly (i.e., no construct-irrelevant variance).
- There is only one correct answer for multiple-choice items.
- No clueing of the answer exists within the item.
- Each item is text dependent; in other words, all information needed to answer the question successfully is contained within the item (or passage).

If the item met the above criteria, the assessment specialist continued the first content review. Using a review checklist, the assessment specialist reviewed the item for content accuracy, including compliance with the content limits in the item specifications, logical and plausible distractors and rationales, accuracy of facts and sources, and language and grade appropriateness. The assessment specialist also reviewed any art requests.

At this point, the items received a second internal review and update. Next, once any approved art was attached to the items, the items were submitted to the copyeditors, who reviewed items and the accompanying art for correct spelling, grammar, and adherence to the style guide. Copyeditors used a checklist that covers context, consistency, and accuracy to guide this review. If any content issues were found during copyediting, the item was returned to the assessment specialist for consideration and correction.

2.4. Content and Bias Review

The purpose of the Content and Bias Review (CBR) meeting is to have Maine educators evaluate new test items developed for the field test item bank. Educators review content, alignment to standards, and the key for all items with the goal of gaining actionable feedback on all items. Only the grade 10 items went through a Content and Bias Review meeting for 2023, since this was the only grade developed specifically for Maine this year.

In Maine, a pre-meeting review took place so that educators could review the items prior to the CBR meeting (training slides can be found in Appendix I). The CBR meeting begins with a general session in which participants are given an overview of the purpose of the meeting and the process to be followed. Training takes place on the criteria that is to be used to evaluate items. Following the general session, participants report to either the reading or math breakout room.

Each breakout room includes a facilitator, and each participant uses a computer connected to the internet in order to access the items via the online review portal. The facilitator provides a brief training on how to view items, as well as how to make comments and judgments in the system (selecting “accept,” “accept with revisions,” or “reject” for each item).

Following the training, items and comments are reviewed individually (organized by passage sets in reading) by the educators. Any items with comments are displayed to the group, and the facilitator leads a discussion regarding any required revisions that are then reconciled with Maine DOE in the days following the meeting, prior to revisions being applied.

Educators review items and provide comments based on the following criteria that is provided on the checklists.

- Items:
 - Item aligns to the standards.
 - Item is clearly worded.
 - Item has one and only one correct answer.
 - Item is mathematically correct.

PDF copies of the Achievement Level Descriptors and the item review criteria checklist are available for the educators to use during their review.

Table 2.11 outlines the total number of items taken to the Content and Bias Review meeting, as well as the number of items accepted, accepted with revisions, and rejected.

Table 2.11. 2023 Content and Bias Review Results

Content Area	Total Items Reviewed	Accepted	Accepted with Revisions	Rejected
Reading	120	90	29	1
Math	145	127	18	0

2.5. Data Review

Data review is an important step for all new item development. It allows for a close examination of item performance and gives an opportunity to remove poorly performing items from the pool or to revise and re-field test poorly performing items.

NWEA adheres to the *Standards for Educational and Psychological Testing* by implementing quality control procedures to ensure accurate information about student learning. The requirements regarding test administration, scoring, and reporting are as follows:

- Standard 4.8: The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria.
- Standard 6.0: Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected.
- Standard 6.9: Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected (AERA et al., 2014).

A data review took place in October/November 2023. Field test items were flagged based on statistical criteria. NWEA assessment specialists then conducted a close examination of the items based on the flags. As a result, some items were removed from the pool, some were deemed appropriate to remain in the pool and changed to an operational status, and some were revised and will be re-field tested in Spring 2024. Table 2.12 presents the criteria for flagging items that were field tested in Spring 2023.

Table 2.12. Data Review Flagging Criteria—Multiple-Choice and Non-Multiple-Choice Items

Type	Label	Statistic	Flag
MC items	Pvalue_LOW/ Pvalue_HIGH	<i>P</i> value	< 0.2 or > 0.9
	Pvalue_Dis	Option percentages	Distractor % > <i>p</i> value
	Pbis_LOW	Item-total correlation	< 0.20
	Pbis_Dis	Item-total correlation for distractors	> 0.05
Non-MC items (Both 1- and 2-point items)	Pvalue_LOW/ Pvalue_HIGH	<i>P</i> value	< 0.2 or > 0.9
	N_012	Low student count for each score	= 0
	Pbis_LOW	Item-total correlation	< 0.2
	Score_0_Pbis	Item-total correlation for score of 0	> 0.0
	Score_0Vs1_Pbis	Item-total correlation for score of 0 > item-total correlation for score of 1	
Non-MC items (2-point items only)	Score_1Vs2_Pbis	Item-total correlation for score of 1 > item-total correlation for score of 2	
	Score_2_Pbis	Item-total correlation for score of 2	< 0.2
Item Parameters	itemFlag_IRT_Parameter	IRT difficulty or step parameters are extreme	≥ 4.25
	itemFlag_IRT_ReversedStep	Reversed step parameters	Step 1 > Step 2
DIF	itemFlag_Gender_DIF/ itemFlag_Black_DIF/ itemFlag_Hispanic_DIF	DIF of gender or ethnicity	C+ or C-

Table 2.13. Data Review Results (forthcoming)

Data Review is scheduled for October/November 2023, and details will be added after that meeting.

Section 3: Administration and Security

3.1. Administration

District and School Assessment Coordinators are primarily responsible for ensuring a uniform assessment administration, including scheduling logistics, training and supervision of proctors, and maintaining assessment security. *The Maine Through Year Assessment Coordinator Guide* provides clear guidance on preparing for, monitoring, and concluding the administration of the Maine Through Year Assessment. *The Maine Through Year Assessment Administration Guide* contains explicit directions and proctor scripts for consistency of administration across different schools and School Administrative Units (SAUs).

3.2. Spring 2023 Administration

This section provides an overview of the observed demographics of participating students, their estimated ability distributions, and descriptions of the item pool.

3.2.1. Student Population

Table 3.1–Table 3.3 display demographic information and ability distributions for Maine’s general student population.

Table 3.1. Demographic Information—Reading

Grade	Type	Total	Gender		Ethnicity						
			Female	Male	Am. Indian/ AK Native	Asian	African American	Hispanic	Native HI/ Pac. Islander	White	Two or More Races
3	N	12091	5915	6175	415	76	154	502	17	10456	471
	%	100	49	51	3.4	0.6	1.3	4.2	0.1	86.5	3.9
4	N	12060	5842	6216	376	109	161	529	18	10435	432
	%	100	48	52	3.1	0.9	1.3	4.4	0.1	86.5	3.6
5	N	11853	5806	6044	370	94	146	479	15	10304	445
	%	100	49	51	3.1	0.8	1.2	4.0	0.1	86.9	3.8
6	N	12041	5944	6094	343	88	165	537	5	10475	428
	%	100	49	51	2.8	0.7	1.4	4.5	0.0	87.0	3.6
7	N	12188	5829	6356	377	89	150	542	16	10599	415
	%	100	48	52	3.1	0.7	1.2	4.4	0.1	87.0	3.4
8	N	12581	6092	6484	381	109	185	540	19	10914	433
	%	100	48	52	3.0	0.9	1.5	4.3	0.2	86.7	3.4
HS	N	12158	5914	6240	413	81	226	497	16	10557	368
	%	100	49	51	3.4	0.7	1.9	4.1	0.1	86.8	3.0

Table 3.2. Demographic Information—Mathematics

Grade	Type	Total	Gender		Ethnicity						
			Female	Male	Am. Indian/ AK Native	Asian	African American	Hispanic	Native HI/ Pac. Islander	White	Two or More Races
3	N	12151	5948	6202	423.0	76.0	155.0	547.0	17.0	10461.0	472.0
	%	100	49	51	3.5	0.6	1.3	4.5	0.1	86.1	3.9
4	N	12138	5872	6264	386.0	109.0	165.0	574.0	18.0	10450.0	436.0
	%	100	48	52	3.2	0.9	1.4	4.7	0.1	86.1	3.6
5	N	11919	5833	6082	375.0	94.0	149.0	530.0	15.0	10311.0	445.0
	%	100	49	51	3.1	0.8	1.3	4.4	0.1	86.5	3.7
6	N	12080	5965	6112	349.0	88.0	166.0	570.0	5.0	10475.0	427.0
	%	100	49	51	2.9	0.7	1.4	4.7	0.0	86.7	3.5
7	N	12253	5860	6390	384.0	91.0	151.0	576.0	15.0	10624.0	412.0
	%	100	48	52	3.1	0.7	1.2	4.7	0.1	86.7	3.4

Grade	Type	Total	Gender		Ethnicity						
			Female	Male	Am. Indian/ AK Native	Asian	African American	Hispanic	Native HI/ Pac. Islander	White	Two or More Races
8	N	12625	6110	6510	388.0	110.0	188.0	576.0	19.0	10912.0	432.0
	%	100	48	52	3.1	0.9	1.5	4.6	0.2	86.4	3.4
HS	N	12192	5939	6249	417.0	85.0	229.0	544.0	15.0	10532.0	370.0
	%	100	49	51	3.4	0.7	1.9	4.5	0.1	86.4	3.0

Table 3.3. Ability Distribution

Grade	Summative Theta			
	Reading		Mathematics	
	Mean	SD	Mean	SD
3	-0.27	1.46	-0.48	1.66
4	0.20	1.38	-0.33	1.78
5	0.14	1.32	-0.12	1.73
6	0.08	1.25	-0.33	1.73
7	0.18	1.32	-0.72	1.74
8	0.43	1.29	-0.67	1.52
HS	0.25	1.00	-1.36	0.87

3.2.2. Item Pool Characteristics

To ensure the adequacy of the item pool for administering a computer adaptive test (CAT), Table 3.4 details the number of items of various types and levels in the item pool for Maine by content category in the summative item pools for reading and mathematics. High school students were administered a 30-item fixed form for the spring summative assessment. Items included in the fixed forms were operational field test items.

Table 3.4. Number of Items by Content Category (Spring 2023 Summative Item Pool)

Content Area	Content Category	Grade						
		3	4	5	6	7	8	HS
Reading	Informational Text	266	134	134	188	226	221	14
	Literary Text	252	183	156	177	204	212	9
	Vocabulary	97	142	93	96	88	97	7
	Total	615	459	383	461	518	530	30
Mathematics	Geometry	20	79	70	74	133	170	8
	Measurement and Data	242	111	84	–	–	–	–
	Numbers and Operations	188	340	310	–	–	–	–
	Operations and Algebraic Thinking	187	99	75	245	174	231	14
	Statistics and Probability	–	–	–	89	201	86	4
	The Real and Complex Number Systems	–	–	–	270	194	26	4
	Total	637	629	539	678	702	513	30

Beyond the content categories, the lower standard levels were also examined by assessing the number of items available at each standard. The percentages of students who received at least one item from each standard are shown in Appendix B.

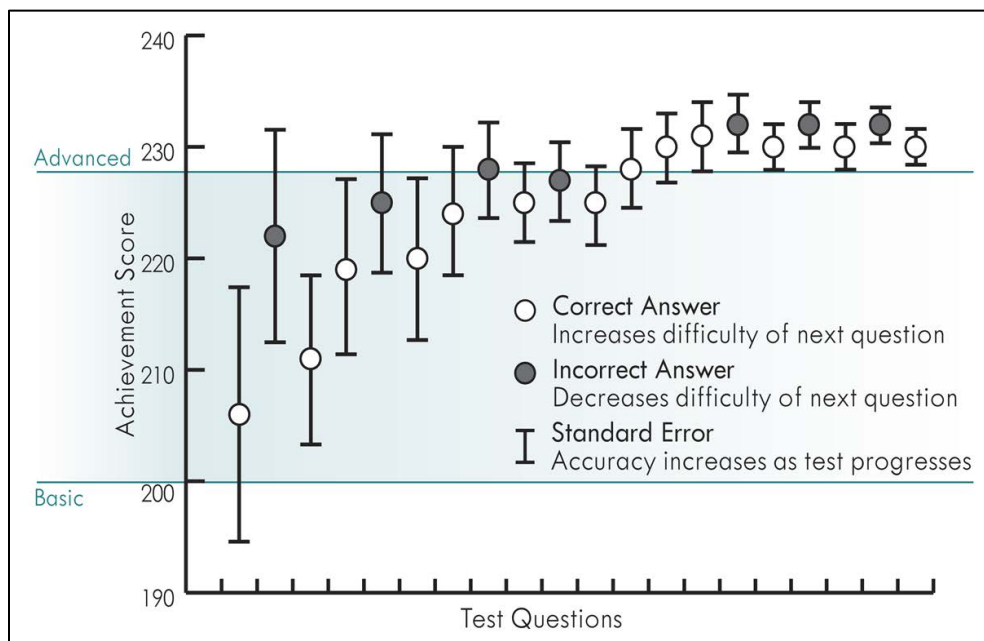
3.3. Constraint-Based Adaptive Test Engine

A CAT administers items to match the ability level of the students: students receive different items based on item difficulty and their ability levels. For example, students with lower ability

levels (based on their answers to previous items) receive easier items compared with students with higher ability levels who receive harder items as the test progresses.

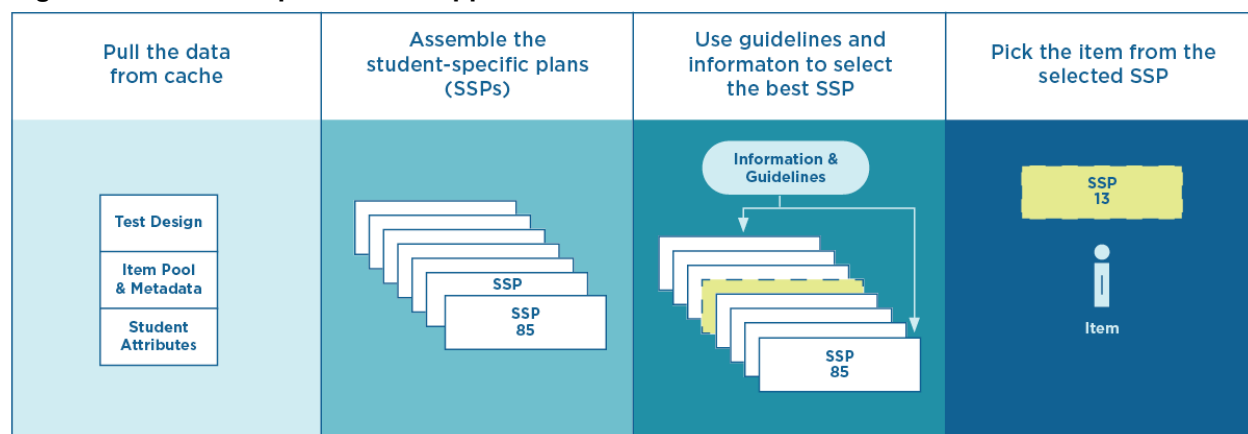
The constraint-based engine (CBE) uses the blueprint and a student’s momentary theta (θ) to drive item selection, as shown in Figure 3.1. Momentary theta is the ability estimate of the student that is recalculated and updated after answering each item. Items are selected based on item difficulty. The goal of the constraint-based engine’s item selection is to provide a test that meets “must-have” constraints and “nice-to-have” guidelines. For example, a constraint of the summative portion is that the engine must deliver 70% on-grade items, while the remaining 30% may adapt by one grade level below or above. The CBE has two stages of consideration as it selects the items necessary to conform to the test blueprint while providing the maximum information about the student based on the student’s momentary ability estimate.

Figure 3.1. Adaptive Engine Overview



The student-specific plan (SSP), similar to the shadow test approach (Van der Linden & Reese, 1998), selects items based on the required aspects of the test blueprint and the student’s momentary theta, as shown in Figure 3.2. Item selection for the SSP occurs through a process of choosing multiple feasible SSPs and then choosing the complete SSP that best maximizes guideline adherence and information. Only after the best SSP has been chosen are items ordered (NWEA, 2020).

Figure 3.2. Student-Specific Plan Approach



Note. Selections are based on the similar shadow test approach.

3.3.1. Engine Evaluation

NWEA checks the adaptive engine at two points: pre-administration simulations and post-administration evaluation. These two studies are important evidence, along with post-administration analyses, for confirming interpretation and test-score use arguments regarding student proficiency with the state standards.

Pre-administration simulations were conducted prior to the operational testing window to evaluate the CBE’s item-selection algorithm and estimation of student ability based on the test blueprints and adaptive specifications. The simulation tool used the operational CBE, thereby providing results with the same properties and functionality as what would be seen operationally. Detailed information regarding the simulation study can be found in Appendix C. After the testing window closed, a post-administration evaluation study was conducted to determine whether the CBE performed as expected. The results of the post-administration evaluation study are presented in this section.

In order to deliver a quality test, various constraints and guidelines are set up in the CBE to specify details of the test requirements. While constraints are rules that must be followed, weights are used to differentiate the importance of different guidelines. One constraint is meeting the requirements of the test blueprint. Because the adaptive test selects items according to individual student abilities in order to provide reliable scores, score precision and item-exposure rate are also important factors. Results for blueprint constraint accuracy, item-exposure rates, and score precision and accuracy are presented below.

3.3.2. Blueprint Constraint Accuracy

Table 3.5 presents the blueprint constraint results at the reporting category level of the spring administration. This analysis exclusively focused on students who completed the maximum/full-length test for each test event, and, in all cases, it yielded a perfect match for the number of items at the reporting category level.

Table 3.5. Blueprint Constraint Accuracy by Reporting Category

Grade	Summative Content Across Instructional Areas	#Items Intended		#Items Administered			%Match
		Min	Max	Average	Min	Max	
Reading							
3	Literary Text	12	14	12	12	14	100
	Informational Text	8	9	9	8	9	100
	Vocabulary	5	7	6	5	7	100
4	Literary Text	11	12	13	11	12	100
	Informational Text	9	11	10	9	11	100
	Vocabulary	5	7	5	5	7	100
5	Literary Text	9	11	11	9	11	100
	Informational Text	9	11	10	9	11	100
	Vocabulary	5	7	6	5	7	100
6	Literary Text	9	11	11	9	11	100
	Informational Text	11	12	11	11	12	100
	Vocabulary	5	7	5	5	7	100
7	Literary Text	8	9	10	8	9	100
	Informational Text	12	14	12	12	14	100
	Vocabulary	5	7	5	5	7	100
8	Literary Text	8	9	11	8	9	100
	Informational Text	12	14	12	12	14	100
	Vocabulary	5	7	4	5	7	100
HS	Literary Text	8	9	9	9	9	100
	Informational Text	12	14	14	14	14	100
	Vocabulary	5	7	7	7	7	100
Mathematics							
3	Operations and Algebraic Thinking	6	6	6	6	6	100
	Numbers and Operations	9	9	9	9	9	100
	Measurement and Data	8	8	8	8	8	100
	Geometry	4	4	4	4	4	100
4	Operations and Algebraic Thinking	5	5	5	5	5	100
	Numbers and Operations	13	13	13	13	13	100
	Measurement and Data	5	5	5	5	5	100
	Geometry	4	4	4	4	4	100
5	Operations and Algebraic Thinking	4	4	4	4	4	100
	Numbers and Operations	14	14	14	14	14	100
	Measurement and Data	5	5	5	4	5	100
	Geometry	4	4	4	4	4	100

Grade	Summative Content Across Instructional Areas	#Items Intended		#Items Administered			%Match
		Min	Max	Average	Min	Max	
6	Operations and Algebraic Thinking	7	7	7	7	7	100
	The Real and Complex Number Systems	12	12	12	12	12	100
	Geometry	4	4	4	4	4	100
	Statistics and Probability	4	4	4	4	4	100
7	Operations and Algebraic Thinking	5	5	5	5	5	100
	The Real and Complex Number Systems	11	11	11	11	11	100
	Geometry	6	6	6	6	6	100
	Statistics and Probability	5	5	5	5	5	100
8	Operations and Algebraic Thinking	13	13	13	13	13	100
	The Real and Complex Number Systems	4	4	4	4	4	100
	Geometry	6	6	6	6	6	100
	Statistics and Probability	4	4	4	4	4	100
HS	Operations and Algebraic Thinking	14	14	14	14	14	100
	The Real and Complex Number Systems	4	4	4	4	4	100
	Geometry	8	8	8	8	8	100
	Statistics and Probability	4	4	4	4	4	100

3.3.3. Score Precision

Conditional standard error of measurement (CSEM) quantifies the degree of measurement error in scale score units, and its calculation is contingent on the student's ability. This means that the test exhibits varying levels of error at different positions along the ability scale. In the context of an adaptive assessment, the CSEM will vary for identical scale scores. Therefore, it is imperative to provide averages in reporting.

In the context of item response theory (IRT), CSEMs for each scale score are defined as the reciprocal of the square root of the test information function (Hambleton & Swaminathan, 1985).

$$CSEM(\theta) = \frac{1}{\sqrt{I(\theta)}},$$

where $CSEM(\theta)$ is the IRT CSEM for a scale score, and $I(\theta)$ is the test information function. CSEMs are especially useful for characterizing measurement precision with respect to score thresholds employed in decision-making, such as the cut score used to determine student proficiency on an assessment. Table 3.6 presents the CSEMs for the achievement level cut scores that demark the three cut scores on the Maine Through Year Assessment. It includes data on the number of students within ± 10 scale score points from these thresholds, the mean

CSEMs for students in proximity to the cut scores, and the standard deviation (SD) of the CSEMs. In general, CSEMs of middle-range scale scores and cut scores are smaller than those at the two ends, indicating low measurement error and high score precision.

Table 3.6. CSEMs at the Cut Scores

Content Area	Grade	<i>Below State Expectations</i>			<i>At State Expectations</i>			<i>Above State Expectations</i>		
		N	Mean CSEM	SD	N	Mean CSEM	SD	N	Mean CSEM	SD
Reading	3	2979	5.1	0.4	4236	5.0	0.2	3047	5.1	0.4
	4	3190	5.0	0.3	3283	5.0	0.2	3676	5.1	0.3
	5	2860	5.0	0.3	2775	5.0	0.7	3628	5.1	0.7
	6	2922	5.1	0.4	3680	5.0	0.3	3050	5.1	0.5
	7	2932	5.1	0.5	4037	5.0	0.5	3045	5.0	0.3
	8	2860	5.0	0.4	4601	5.0	0.4	3431	5.0	0.5
	HS	4284	5.4	1.1	2864	5.0	0.6	2959	6.3	1.0
Mathematics	3	3130	5.0	0.2	3381	5.0	0.1	2923	5.0	0.1
	4	4125	4.0	0.2	3053	4.0	0.1	2819	4.0	0.2
	5	3875	4.0	0.2	4072	4.0	0.1	2384	4.1	0.2
	6	3640	5.0	0.1	3640	5.0	0.1	1767	5.0	0.2
	7	4184	4.0	0.3	4184	4.0	0.2	1842	4.0	0.2
	8	5074	4.0	0.5	4299	4.0	0.1	1617	4.0	0.2
	HS	6608	7.8	1.1	2811	7.0	0.1	1334	7.1	1.9

Table 3.7 presents the average CSEM by score decile, including the overall student ability distribution. A decile is similar to a percentile rank, with 10 ranks corresponding to the 10th, 20th, 30th, . . . 90th, and 100th percentile ranks. A higher SEM indicates a shallower pool of items suitable for students with these abilities. For instance, results indicate that the summative reading item pool is notably limited for students with very high abilities, while the mathematics item pool is shallower for students with very low and high abilities.

Table 3.7. CSEMs by Score Decile

Grade	Overall	Proficiency Score Decile									
		1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Reading											
3	4.1	4.4	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.1	4.5
4	4.1	4.3	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.6
5	4.1	4.2	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.1	4.7
6	4.1	4.4	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.6
7	4.1	4.2	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.4
8	4.1	4.3	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.4
HS	3.6	4.2	3.4	3.0	3.0	3.0	3.0	3.1	4.0	4.0	5.3
Mathematics											
3	4.1	4.3	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.4
4	4.0	4.1	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.2
5	4.1	4.1	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.3

Grade	Overall	Proficiency Score Decile										
		1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
6	4.0	4.2	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.2
7	4.1	4.2	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.1
8	4.1	4.3	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.1
HS	4.4	5.4	5.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.3

3.3.4. Item-Exposure Rates

Because different students receive different items based on blueprint constraints and their ability during an adaptive administration, it is ideal to have a low exposure rate. However, being the first administration of the Through Year Assessment, post-equating is required to derive the Maine scale. It is necessary to secure an adequate item-exposure rate for most of the items. Thus, the item pool was reduced to fulfill this plan.

The exposure rate for each operational item was calculated as the percentage of students who received that item, as shown in Table 3.8. For example, if Item 1 was administered to 500 out of 1,000 students, the exposure rate would be 50%. “Total” is the total number of items in the operational item pool. For fixed forms administered in high school, a 100% exposure rate is anticipated. However, an exposure rate of only 81–99% was observed due to incomplete test events.

Table 3.8. Operational Item Exposure Rates

Grade	#Items				Item Exposure Rate											
	Total	Used	Unused	Unused %	0–20%		21–40%		41–60%		61–80%		81–99%		100%	
					N	%	N	%	N	%	N	%	N	%	N	%
Reading																
3	1074	591	483	45.0	567	95.9	21	3.6	3	0.5	0	0.0	0	0	0	0
4	1457	849	608	41.7	829	97.6	18	2.1	2	0.2	0	0.0	0	0	0	0
5	1303	783	520	39.9	766	97.8	9	1.1	5	0.6	3	0.4	0	0	0	0
6	1362	788	574	42.1	765	97.1	20	2.5	3	0.4	0	0.0	0	0	0	0
7	1509	807	702	46.5	790	97.9	17	2.1	0	0.0	0	0.0	0	0	0	0
8	1048	553	495	47.2	540	97.6	11	2.0	2	0.4	0	0.0	0	0	0	0
HS	30	30	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	100	100	0	0
Mathematics																
3	1266	647	620	49.0	636	98.3	8	1.2	2	0.3	1	0.2	0	0	0	0
4	1805	960	847	46.9	956	99.6	4	0.4	0	0.0	0	0.0	0	0	0	0
5	1846	644	1204	65.2	639	99.2	4	0.6	1	0.2	0	0.0	0	0	0	0
6	1919	682	1239	64.6	678	99.4	4	0.6	0	0.0	0	0.0	0	0	0	0
7	1893	668	1226	64.8	652	97.6	16	2.4	0	0.0	0	0.0	0	0	0	0
8	1215	529	686	56.5	518	97.9	10	1.9	1	0.2	0	0.0	0	0	0	0
HS	30	30	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	100	100	0	0

A number of field test items were embedded in the Spring 2023 test for possible operational use in future test administrations. Field test items were distributed using target demographic characteristics of the Maine student population. For example, each item should be administered to approximately 50% female and 50% male students if the Maine student population has a 50/50 gender proportion. The results presented in Table 3.9 show that all field test items were appropriately administered to each demographic subgroup.

Table 3.9. Field Test Item Exposure Rates

Grade	FT Items	Mean	SD	Female	Male	African Indian	Asian	African American	Hispanic	Native Hawaiian or Pacific Islander	Caucasian	Two or More Races
Reading												
3	60	1405.9	1549.4	49.5	50.5	3.6	0.6	1.3	4.6	0.2	85.6	4.2
4	126	669.0	305.9	49.0	50.9	3.1	1.0	1.3	4.5	0.2	86.4	3.5
5	98	845.1	450.7	49.3	50.7	3.2	0.8	1.3	4.3	0.3	86.7	3.7
6	87	965.3	872.1	49.7	50.3	2.6	0.8	1.3	4.0	0.2	87.8	3.5
7	137	618.0	343.9	48.7	51.3	3.1	0.8	1.2	4.7	0.3	86.8	3.4
8	130	675.0	295.1	48.7	51.3	3.0	0.8	1.4	4.2	0.3	87.0	3.4
HS	19	3168.5	4468.5	49.1	50.9	3.5	0.9	2.3	6.1	0.1	84.7	2.5
Mathematics												
3	160	530.6	92.0	49.2	50.8	3.5	0.6	1.3	4.5	0.3	86.1	3.9
4	153	555.0	140.6	48.8	51.2	3.2	0.9	1.4	4.7	0.3	86.1	3.6
5	160	520.8	43.1	49.1	50.9	3.1	0.8	1.2	4.4	0.2	86.6	3.7
6	160	527.6	67.0	49.5	50.5	2.9	0.7	1.4	4.7	0.2	86.7	3.5
7	159	537.7	138.4	48.6	51.4	3.1	0.8	1.2	4.7	0.3	86.7	3.4
8	159	554.2	154.5	49.0	51.0	3.0	0.9	1.5	4.5	0.3	86.5	3.4
HS	38	1594.1	999.4	49.2	50.8	3.4	0.6	2.0	5.0	0.2	85.8	3.0

3.3.5. Item Sequence

The distribution of items that each student receives is not inherently subject to a predefined sequence or grouping for the summative, diagnostic (MAP Growth), and field test items. In the absence of specific preferences, the adaptive engine arranges the items based on the individual student’s test performance. An exception to this rule pertains to items that are part of a set with a common reading passage or paired passages; in such cases, the engine ensures that these items are delivered as a cohesive group rather than being dispersed. NWEA’s evaluation reveals that items were allocated based on their performance without adhering to any predefined sequence or grouping, except for the designated locations for the field test items. In the reading tests, the actual placement of field test items varied due to the arrangement of reading passage sets and the engine’s design to avoid introducing unrelated items in the midst of a reading passage set.

3.4. Paper Form Administration

For the Spring 2023 assessment, the majority of Maine’s students participated through the computer adaptive assessment. Students with an IEP or 504 Plan could request an alternate, accommodated paper-based form in standard print, large print, or braille. A fixed test form was built for each grade and content area to fulfill the needs of the three accommodated test forms. Braille forms were prepared in advance according to registration data, and the required materials were packed and shipped to the requesting schools. Standard and large print paper-based forms were available via print on demand. These materials were sent to School Assessment Coordinators via NWEA’s secure SFTP site.

Table 3.10 presents the number of summative operational items needed for the spring fixed forms.

- All items are on grade level.
- There are no anchor or linking items on the paper forms.

Table 3.10. Paper Form Summative Item Totals by Content and Grade

Content	Grade	Summative Operational
Reading	3–8	27
Math	3–8	27
Reading	10	30
Math	10	30

3.5. Spring 2023 Fixed Form Blueprints

Table 3.11–Table 3.15 display the item counts by instructional area for the Spring 2023 assessments.

Table 3.11. Reading Item Counts by Instructional Area, Grades 3–8

Instructional Area	Total						Summative						Diagnostic (MAP Growth)					
	G3	G4	G5	G6	G7	G8	G3	G4	G5	G6	G7	G8	G3	G4	G5	G6	G7	G8
Literary Text	16	15	15	14	13	13	12	11	11	9	8	8	4	4	4	5	5	5
Informational Text	13	14	14	15	16	16	8	9	9	11	12	12	5	5	5	4	4	4
Vocabulary	12	12	12	12	12	12	7	7	7	7	7	7	5	5	5	5	5	5
Total Items	41	41	41	41	41	41	27	27	27	27	27	27	14	14	14	14	14	14

Table 3.12. Reading Item Counts by Instructional Area, Grade 10

Instructional Area	Total	Summative	Diagnostic
	G10	G10	G10
Literary Text	13	9	4
Informational Text	17	13	4
Vocabulary	12	8	4
Total Items	42	30	12

Table 3.13. Mathematics Item Counts by Instructional Area, Grades 3–5

Instructional Area	Total			Summative			Diagnostic		
	G3	G4	G5	G3	G4	G5	G3	G4	G5
Operations and Algebraic Thinking	10	10	9	6	5	4	4	5	5
Numbers and Operations	13	17	18	9	13	14	4	4	4
Measurement and Data	12	9	9	8	5	5	4	4	4
Geometry	10	9	9	4	4	4	6	5	5
Total Items	45	45	45	27	27	27	18	18	18

Table 3.14. Mathematics Item Counts by Instructional Area, Grades 6–8

Instructional Area	Total			Summative			Diagnostic		
	G6	G7	G8	G6	G7	G8	G6	G7	G8
Operations and Algebraic Thinking	11	10	17	7	5	13	4	5	4
The Real and Complex Number Systems	16	15	9	12	11	4	4	4	5
Geometry	9	10	10	4	6	6	5	4	4
Statistics and Probability	9	10	9	4	5	4	5	5	5
Total Items	45	45	45	27	27	27	18	18	18

Table 3.15. Mathematics Item Counts by Instructional Area, Grade 10

Instructional Area	Total	Summative	Diagnostic
	G10	G10	G10
Operations and Algebraic Thinking	17	13	4
The Real and Complex Number Systems	9	5	4
Geometry	12	8	4
Statistics and Probability	9	4	5
Total Items	47	30	17

3.5.1. Receive and Take Inventory of School Materials

The quantity of materials shipped to each school is based on data collected during the rostering process. School Assessment Coordinators are required to open packages containing braille forms immediately upon receipt to inventory the contents. School Assessment Coordinators are responsible for the printing and secure handling of standard and large print paper-based forms, as well as for providing secure assessment materials to proctors. All standard assessment booklets are provided as single materials. School Assessment Coordinators do not distribute any assessment materials, except the *Maine Through Year Assessment Proctor User Guide* and *The Maine Through Year Assessment Administration Guide*, until the day of each session.

On the day of the assessment, the School Assessment Coordinator distributes the correct assessment booklets needed for that day's assessment to each proctor. Assessment booklets are distributed to proctors early enough on the day of the assessment to give them ample time to review the directions prior to the assessment. After each day of the assessment is complete, all assessment materials are returned to the School Assessment Coordinator for secure storage as soon as possible. All materials, including used and unused booklets and scratch paper, are returned at the end of each day of testing.

3.5.2. Score Transcription

During or immediately following assessment administration, student responses for paper-based accommodated assessments are transcribed into the online assessment engine. To transcribe responses requires the proctor or other designated and authorized district or school personnel to log in to the NWEA State Solutions Secure Browser using the student's test ticket. The required steps for the proctor to transcribe student answers are as follows:

1. Obtain the student's test ticket from the School Assessment Coordinator.
2. After the student has completed the paper accommodated assessment, use a device that has the NWEA State Solutions Secure Browser software installed and use the student's test ticket to log in to the student's assessment.
3. For security reasons, Maine DOE recommends, when feasible, that a second trained staff member be present to verify all transcriptions.
4. Once transcribing student responses is complete, the assessment is submitted. The proctor should then return all printed assessment materials to the School Assessment Coordinator.

Transcribing is the process of moving the student's assessment response to another medium by a district employee. The process should be as faithfully completed as possible and follow the qualifications and procedures as outlined:

1. The transcriber must be a current employee of the school district.
2. The transcriber must be trained in assessment administration and have signed the Assessment Security and Data Privacy Agreement.
3. Transcription must take place in a secure location.
4. The assessment must be transcribed exactly as the student answered the assessment items.

Local SAU policy will determine whether School Assessment Coordinators should securely destroy test tickets, scrap paper, and accommodated paper forms on-site or if all materials should be sent to the district office to be securely destroyed by the District Assessment Coordinator. If shipping to the district office, security and record-keeping guidance must be followed.

3.6. Assessment Security

In a centralized assessment process, it is critical that equity of opportunity, standardization of procedures, and fairness to students is maintained. Therefore, Maine DOE requires that all assessment administrators and proctors review the information in the *Maine Assessment Security Handbook*.

The Maine DOE recommends that assessment administrators (or proctors) report any potential irregularities to the School Assessment Coordinator. This is especially important for any irregularities that may:

- (1) involve a breach of assessment item security
- (2) lead to assessment invalidation
- (3) involve student misbehavior
- (4) involve educator misbehavior

The School Assessment Coordinator, or other administrator, should report irregularities to Krista Averill, Maine DOE Assessment Coordinator, at Krista.Averill@maine.gov or 1-207-215-6528. See the *Maine Assessment Security Handbook* for more details on this process.

3.6.1. Assessment Ethics and Appropriate Practice

All teachers need to be familiar with appropriate assessment ethics and security practices related to assessments. Proctors are expected to actively monitor student participation during the assessment to ensure students remain on-task. Professionalism, common sense, and practical procedures provide the right framework for assessment ethics. The *Maine Assessment Security Handbook* outlines clear practices for appropriate security.

3.6.2. Online Security

Student test tickets contain student-level password information for accessing the assessment and must be kept secure. Proctors should print or be given the student test tickets prior to assessment administration, allowing them ample time to review and organize the tickets for distribution before the assessment begins. Once an assessment session is started, only the

student taking the assessment is allowed to view the student's screen. No one is allowed to view or copy assessment content while a student is taking the assessment.

The Maine Through Year Assessment Coordinator Guide, as well as other manuals and guides available online, are not considered secure assessment materials.

3.6.3. Student Assessment Security

Students should look only at their individual computers. For further security, folders may be set up around each computer screen to eliminate any possibility of students looking at other computer screens. For larger groups, it is advisable to have a sufficient number of proctors to monitor the room.

3.6.4. Returning or Destroying Secure Materials

Proctors should collect all student test tickets, scratch paper, and assessment booklets (where applicable) from students after the assessment so that those materials can be securely destroyed.

3.7. Systems for Protecting Data Integrity and Privacy

School Assessment Coordinators, assessment administrators, and proctors are required to complete and sign the MEA Assessment Security and Data Privacy Agreement. Signed copies should be filed and kept on-site, available for delivery to the Maine DOE if requested.

NWEA maintains the following protocols to ensure that the sensitive data that is captured are protected and secure from unauthorized use, hacks, or other forms of compromise.

Test Content Security

NWEA encrypts test data both prior to transmission and in-transit and then delivers the data through a secure downloadable browser that is only accessible through 256-bit TLS user authentication and proctor-provided usernames and passwords. NWEA's test system also saves students' work at frequent intervals, and assessment packages are encrypted while on students' workstations.

Data Protection

- Data at rest are protected across a wide range of Amazon Web Services (AWS) and state applications.
- Encryption is enabled for all network traffic, including Transport Layer Security for web-based network infrastructure
- Policies and procedures to protect personally identifiable information (PII) data are strictly enforced.

Secure Identity and Access Management

- A centralized identity provider is used to manage account access, restricting access to authorized personnel only.
- A least privilege model is used to ensure operational staff have only those privileges needed to complete their tasks.
- Multi-factor authentication and other account-level controls are enabled.

- Passwords and other credentials are securely stored using AWS tools that handle encryption, rotation, and access control.

Infrastructure Protection

- Operating systems, middleware, applications, and code are patched on a regular basis.
- Distributed Denial of Service (DDoS) protection layers are used for all internet-facing applications.
- Intrusion detection/prevention services are utilized.
- Inbound and outbound traffic is controlled and monitored based on established rules.

Detection and Monitoring

- AWS are leveraged to comprehensively monitor all layers.
- Application and system-level logs are analyzed on a periodic basis to gain insights into the information contained in these logs.
- An incident management process is maintained for security events that may affect the confidentiality, integrity, or availability of systems or data.
- Monitoring and alerts are configured and investigated on a regular basis for any events that are unexpected or do not make sense, including hacking attempts and attacks.

Section 4: Item Statistics, Calibration, and Scaling

Being the first administration of the Maine Through Year Assessment, a new Maine scale and achievement level cut scores have been established using the summative items. Maine DOE, following the Technical Advisory Committee's (TAC) recommendations, determined the Maine scale to be grade-level-specific scales. This section presents item statistics and the methods and process of establishing the Maine scale.

4.1. Classical Item Statistics

4.1.1. Expected *P* Value

Item difficulty is measured by a *p* value that represents the proportion of students who answered an item correctly and ranges from 0 to 1. A high *p* value indicates an easy item, with a high percentage of students answering it correctly, whereas a low *p* value indicates a difficult item. For example, a *p* value of 0.79 indicates that 79% of students answered the item correctly. In the case of polytomous items, the *p* value is calculated as the average item score divided by the number of possible score points on the item.

Table 4.1 and Table 4.2 present the summary statistics for the *p* values across operational and field test items and the count of items falling within different *p*-value ranges (e.g., less than or equal to 0.1, 0.2, etc.). The data include adaptive items for grades 3–8 and fixed-form items for the high school grade. For adaptive items that were administered without a representative student sample, their expected *p* values are provided. An expected *p* value represents the proportion of correct responses if the item was administered to a representative student sample. Appendix D provides the summary *p*-value statistics by item type.

Table 4.1. Summary of P Values—Operational Items

Grade	N	P Value Summary					P Value Counts									
		Mean	Median	SD	Min	Max	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
Reading																
3	591	0.45	0.45	0.18	0.03	0.88	12	39	91	94	106	124	69	43	13	0
4	849	0.51	0.51	0.19	0.04	0.93	12	34	83	122	157	137	139	113	47	5
5	783	0.52	0.52	0.19	0.02	0.91	6	29	63	101	153	150	126	100	53	2
6	788	0.48	0.48	0.18	0.06	0.90	8	46	87	131	141	152	115	83	25	0
7	807	0.53	0.55	0.18	0.03	0.91	4	38	56	92	145	161	145	113	49	4
8	553	0.59	0.61	0.19	0.03	0.97	1	16	29	50	71	98	117	101	64	6
HS	30	0.56	0.56	0.15	0.18	0.88	0	1	1	2	6	7	8	4	1	0
Mathematics																
3	647	0.43	0.42	0.24	0.00	0.96	63	85	70	93	85	78	68	56	42	7
4	960	0.49	0.49	0.25	0.00	0.97	65	87	109	107	125	108	116	111	112	20
5	644	0.48	0.49	0.24	0.00	0.98	47	59	63	90	74	86	84	72	54	15
6	682	0.43	0.41	0.24	0.01	0.96	46	100	89	100	86	79	65	75	35	7
7	668	0.41	0.39	0.24	0.01	0.96	74	93	74	105	77	77	57	57	47	7
8	529	0.40	0.39	0.20	0.03	0.90	31	63	90	92	96	67	43	35	11	1
HS	30	0.29	0.29	0.13	0.04	0.53	3	6	7	7	6	1	0	0	0	0

Note. Items included for the high school grade were operational field test items in the spring administration.

Table 4.2. Summary of P Values—Field Test Items

Grade	N	P Value Summary					P Value Counts									
		Mean	Median	SD	Min	Max	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
Reading																
3	60	0.46	0.47	0.13	0.21	0.74	0	0	9	7	20	17	4	3	0	0
4	126	0.45	0.44	0.15	0.05	0.79	1	4	15	31	33	23	10	9	0	0
5	98	0.47	0.48	0.13	0.17	0.87	0	1	7	22	27	27	10	3	1	0
6	87	0.46	0.48	0.15	0.18	0.82	0	2	14	14	20	25	9	2	1	0
7	137	0.48	0.49	0.14	0.17	0.77	0	1	18	18	39	29	21	11	0	0
8	130	0.50	0.50	0.17	0.10	0.88	0	3	14	19	29	26	23	11	5	0
HS	19	0.51	0.54	0.13	0.27	0.78	0	0	1	3	5	6	2	2	0	0
Mathematics																
3	160	0.41	0.38	0.21	0.04	0.92	7	24	25	30	28	17	13	9	5	2
4	153	0.41	0.37	0.21	0.01	0.95	7	17	28	36	17	18	12	9	7	2
5	160	0.37	0.37	0.19	0.02	0.85	8	27	30	24	31	23	7	7	3	0
6	160	0.29	0.27	0.17	0.02	0.88	19	38	35	29	25	6	3	4	1	0
7	159	0.32	0.29	0.20	0.01	0.82	22	29	31	28	17	15	12	4	1	0
8	159	0.32	0.29	0.20	0.02	0.86	21	38	22	28	17	14	12	5	2	0
HS	36	0.22	0.17	0.17	0.02	0.72	10	10	9	3	0	2	1	1	0	0

Note. Items included for the high school grade were operational field test items in the spring administration.

4.1.2. Item Discrimination (Item-Total Correlation)

Item-total correlation describes the relationship between performance on an item and performance on the entire test (test scaled score). Students who perform well on a test are expected to have a higher probability of selecting the right answer to any given item, and students who perform poorly are more likely to select the wrong answer. This means that for a highly discriminating item, students who get the item correct will have a higher test score than students who get the item incorrect. The item-total correlation coefficient ranges between -1.0 and $+1.0$. An item with a high positive item-total correlation discriminates between low-performing and high-performing students better than an item with an item-total correlation near zero. A negative item-total correlation indicates that lower-performing students did better on that item than higher-performing students. However, if an item is either very difficult or very easy, there will be little variation in student responses, as most students would either respond incorrectly or correctly. The resulting item-total correlation for such items is typically low.

Table 4.3 and Table 4.4 present the summary statistics for the item-total correlations across operational and field items, respectively. Instead of using the number-correct raw score, the estimated final scaled score was used to compute the item-total correlations because number-correct scores would not provide much insight into student performance on an adaptive test. For items administered adaptively in grades 3–8, their item-total correlations tend to be lower because these adaptive items were seen by students within a restricted ability range. Additionally, most of the items displaying negative item-total correlations had very few responses (less than 10 student responses). Appendix E provides the summary item-total correlation statistics by item type.

Table 4.3. Summary of Item-Total Correlations—Operational Items

Grade	N	Item-Total Correlation Summary					Item-Total Correlation Counts									
		Mean	Median	SD	Min	Max	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
Reading																
3	591	0.36	0.37	0.14	-0.30	0.99	27	33	105	198	142	58	14	4	3	1
4	849	0.35	0.36	0.17	-1.00	1.00	40	70	144	271	182	51	18	12	5	5
5	783	0.33	0.34	0.15	-1.00	1.00	30	57	174	288	164	36	13	3	0	2
6	788	0.34	0.35	0.17	-1.00	1.00	34	71	143	253	170	62	13	4	0	6
7	807	0.36	0.36	0.17	-1.00	1.00	29	62	151	240	197	61	23	8	4	5
8	553	0.36	0.37	0.15	-1.00	1.00	12	37	96	190	135	40	10	2	3	5
HS	30	0.44	0.44	0.14	-0.11	0.67	1	0	2	6	9	9	3	0	0	0
Mathematics																
3	647	0.33	0.34	0.13	-0.69	0.81	23	49	154	248	123	33	10	4	1	0
4	960	0.34	0.35	0.13	-0.14	0.78	31	81	212	332	213	69	18	4	0	0
5	644	0.36	0.36	0.11	-0.13	0.98	4	27	130	255	175	40	5	0	1	1
6	682	0.36	0.38	0.20	-1.00	1.00	32	30	107	208	165	56	9	8	4	10
7	668	0.35	0.36	0.18	-0.66	1.00	37	51	108	217	137	56	19	8	3	7
8	529	0.35	0.35	0.15	-0.57	0.94	21	20	103	222	94	20	16	7	4	1
HS	30	0.37	0.39	0.13	0.05	0.60	1	2	4	10	7	5	1	0	0	0

Note. Items included for the high school grade were operational field test items in the spring administration.

Table 4.4. Summary of Item-Total Correlations—Field Test Items

Grade	N	Item-Total Correlation Summary					Item-Total Correlation Counts										
		Mean	Median	SD	Min	Max	≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9	
Reading																	
3	60	0.35	0.39	0.13	-0.01	0.54	5	4	4	18	27	2	0	0	0	0	
4	126	0.32	0.33	0.12	-0.02	0.57	7	13	31	43	28	4	0	0	0	0	
5	98	0.31	0.31	0.12	0.00	0.58	5	12	27	31	20	3	0	0	0	0	
6	87	0.32	0.34	0.12	0.04	0.53	1	14	22	22	25	3	0	0	0	0	
7	137	0.33	0.35	0.12	0.02	0.58	5	18	25	44	39	6	0	0	0	0	
8	130	0.33	0.35	0.14	-0.02	0.59	9	15	22	38	34	12	0	0	0	0	
HS	19	0.35	0.41	0.12	0.12	0.57	0	3	3	3	9	1	0	0	0	0	
Mathematics																	
3	160	0.37	0.38	0.12	-0.02	0.62	2	10	33	43	46	24	2	0	0	0	
4	153	0.39	0.40	0.13	-0.12	0.63	7	7	16	44	53	24	2	0	0	0	
5	160	0.39	0.38	0.12	0.01	0.64	2	9	25	48	44	30	2	0	0	0	
6	160	0.37	0.39	0.14	-0.06	0.62	7	9	33	37	47	25	2	0	0	0	
7	159	0.34	0.34	0.15	-0.08	0.65	11	21	29	34	38	20	6	0	0	0	
8	159	0.36	0.39	0.13	-0.07	0.58	8	10	30	36	59	16	0	0	0	0	
HS	37	0.35	0.37	0.13	0.08	0.59	1	3	9	12	7	5	0	0	0	0	

Note. Items included for the high school grade were operational field test items in the spring administration.

4.2. IRT Calibration

When establishing a new scale, the first step is to calibrate items to a standardized scale, then use the calibrated items to derive student scores. The Rasch model (Rasch, 1960, 1980; Wright, 1977) for dichotomous items and the partial-credit model (PCM; Masters, 1982) for polytomous items were used to calibrate items and create the Maine scale. These two models have had a long-standing presence in applied testing programs. For all content areas, item parameter estimations were implemented using WINSTEPS 3.90.2.0 (Linacre, 2015) that used joint maximum likelihood estimation (MLE), as described by Wright (1977) and Masters (1982).

Under the Rasch model, the probability of a student with ability θ responding correctly to item i is as follows, where θ_j and b_i are the person and item parameters, respectively:

$$P(u_{ij} = 1|\theta_j, b_i) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}$$

Under the PCM, the probability of a student with ability θ having a score at the k th level of item i is:

$$P(u_{ij} = k|\theta_i) = \frac{e^{[\sum_{u=1}^k (\theta_j - b_i + d_{iu})]}}{\sum_{v=1}^{m_i} e^{[\sum_{u=1}^k (\theta_j - b_i + d_{iu})]}}$$

where k is the score on the item, m_i is the total number of score categories for the item, d_{iu} is the threshold parameter for the threshold between scores u and $u - 1$, and θ_j and b_i are the person and item parameters, respectively.

Being the first administration of the Through Year Assessment, the free calibration¹ method was used to derive item parameters for the summative items by subject and grade. Table 4.5 presents the summary of IRT item statistics across all operational items.

Table 4.5. Summary of IRT Item Statistics—Operational Items

Grade	#Items	#Parameters	Mean	Median	SD	Min	Max	Range (Max–Min)
Reading								
3	591	666	0.06	0.00	1.17	-2.98	3.95	6.93
4	849	944	0.13	0.14	1.20	-3.21	3.93	7.14
5	783	846	0.01	0.05	1.13	-2.78	4.18	6.96
6	788	861	0.19	0.17	1.08	-2.63	3.49	6.12
7	807	874	-0.03	-0.09	1.12	-2.76	4.59	7.35
8	553	602	-0.10	-0.14	1.13	-3.99	4.50	8.49
HS	30	41	-0.07	-0.05	0.78	-2.07	2.04	4.11
Mathematics								
3	647	706	0.12	0.00	1.87	-4.88	8.05	12.93
4	960	1075	-0.22	-0.28	1.95	-5.21	7.16	12.37
5	644	731	0.08	-0.08	1.88	-5.01	8.07	13.08
6	682	769	0.23	0.21	1.79	-4.57	6.37	10.94

¹ Calibration can be done by itself or combined with equating. The former is referred to as free calibration, and the latter is the anchor/fixed parameter method.

Grade	#Items	#Parameters	Mean	Median	SD	Min	Max	Range (Max–Min)
7	668	747	-0.04	-0.08	1.93	-4.91	5.44	10.35
8	529	588	0.00	-0.08	1.38	-3.63	3.92	7.55
HS	30	34	-0.21	-0.39	0.88	-1.53	2.17	3.70

Note. Items included for the high school grade were operational field test items in the spring administration.

4.3. IRT Model Assumptions

Being one of the item response theory models (IRT), Rasch and PCM models have the same assumptions as other IRT models: local independency, model fit, and unidimensionality (Hambleton & Swaminathan, 1985). These three assumptions are checked to evaluate the appropriateness of using the Rasch and PCM models for the assessment.

4.3.1. Local Independence

Local independence refers to a response to an item that is not affected by other items after removing the contribution of ability measures. The IRT model assumes that the response to an item is only affected by the item's difficulty and student's ability. Local dependence violates this assumption by introducing factors irrelevant to those two factors. Examples of local independence violation are:

- The response to an item depends on the response to a prior item—such as, derive a value from Item A, then use Item A's response to solve Item B's equation. If Item A is answered incorrectly, then the response to Item B must be wrong. Scores on Item B are affected by the answer to Item A, a factor other than item difficulty and student ability.
- Other items on the test give away the answer to Item A—this is referred to as clueing in test development.

When constructing items, each item has a complete concept in itself and does not rely on other items. When selecting items for an adaptive test, item enemy information is incorporated to avoid clueing.

4.3.2. Model Fit

Model fit refers to how well an item fits the calibration model. It is usually a statistical chi-square, representing the difference between the observed score (i.e., actual student responses to items) and the expected score (i.e., what the model predicts students with a certain ability should be getting on items). Individual item fit is evaluated using infit and outfit statistics:

- **Infit:** an information-weighted fit statistic that is more sensitive to unexpected behavior affecting responses to items near the student's ability level
- **Outfit:** an outlier-sensitive fit statistic that is more sensitive to unexpected behavior by persons on items far from the student's ability level

Both infit and outfit provide mean-square fit (MNSQ) statistics. The expected value of MNSQ is 1.0. Summary statistics for the infit and outfit MNSQ statistics are presented in Table 4.6. The fit statistics were computed using response data from on-grade items with a minimum of 500 responses to ensure statistical stability. A cutoff of greater than 2.0 is used for item-fit flagging.

The table shows that all average infit and outfit values are close to 1.0, indicating that items fit reasonably well at their intended grade level. While some grades have cases of item outfit

values greater than 2.0, the majority of such values are within the value of 3.0. These items have less impact on the measurement system because “outfit problems are less of a threat to measurement than infit ones” (Linacre, 2002). The results from the model fit analyses and item statistics will be used to inform future item development. For instance, if items with model fit statistics that fall outside of the acceptable range are found to be relatively easy or difficult, they will be replaced during item development to ensure proper coverage of the student ability scale.

Table 4.6. Summary of Mean-Square Infit and Outfit Statistics

Grade	N	Infit				Outfit			
		Mean	Min	Max	SD	Mean	Min	Max	SD
Reading									
3	129	0.99	0.57	1.27	0.09	1.01	0.63	1.56	0.15
4	89	1.02	0.85	1.51	0.11	1.05	0.78	2.41	0.20
5	75	1.01	0.87	1.64	0.12	1.03	0.71	2.30	0.22
6	111	1.02	0.87	1.53	0.11	1.05	0.79	1.85	0.20
7	144	1.00	0.75	1.34	0.10	1.02	0.67	1.88	0.19
8	157	1.00	0.84	1.34	0.09	1.03	0.80	2.16	0.17
HS	30	1.00	0.82	1.39	0.12	1.01	0.59	2.24	0.28
Mathematics									
3	254	0.99	0.81	1.55	0.09	1.01	0.68	2.05	0.17
4	223	1.00	0.86	1.74	0.09	1.06	0.79	3.46	0.29
5	184	0.98	0.83	1.24	0.07	1.02	0.81	2.81	0.23
6	199	1.00	0.83	1.44	0.10	1.09	0.78	3.38	0.35
7	202	1.00	0.87	1.64	0.09	1.06	0.81	4.28	0.36
8	241	1.00	0.84	1.58	0.10	1.04	0.80	2.76	0.25
HS	30	0.99	0.76	1.28	0.13	0.97	0.58	1.37	0.20

Note. Items included for the high school grade were operational field test items in the spring administration.

4.3.3. Unidimensionality

The unidimensionality assumption is that items on the test measured only one latent trait. It can be assessed by examining the model fit. Essentially, if the model fit is not adequate, then the unidimensional assumption is not tenable. The specific steps taken and criteria to assess model fit are discussed in detail in the previous section. The results indicate that the unidimensionality assumption holds for most tests.

4.4. Scaling

A scale can be established through different methods (Kolen & Brennan, 2004). The fix two cut score method was selected because it eases the use and interpretation of score and achievement levels. This list shows the steps for implementing this method:

1. Maine DOE determines:
 - a. the number of achievement levels,
 - b. the initial scale score range, and
 - c. two fixed cut scores across grades and content areas.
2. Cut scores are obtained from the standard setting meeting. Note that the recommended cut scores are approved by the school board.
3. The equations below are used to derive equating constants.

4. The lowest and highest obtainable scores (LOSS & HOSS) of the scale are finalized.

Puhan & Dorans (2018) was consulted when determining the scale properties. Relevant key points considered were:

1. The mean score centers around the midpoint of the scale in order to maximize the longevity of the scale.
 - a. Because the fix two cut scores method is used, the *At State Expectations* level cut score should be centered around the midpoint of the scale.
2. The range of scores is wide enough to accommodate population shift. In other words, the number of score units preserves the score differentiation but does not yield unjustified differentiation.
 - a. Puhan and Dorans (2018) recommends that the number of scale units is similar to the raw score points. However, empirical data shows that this approach may cause many scale scores to be rounded to the same values or truncated to LOSS/HOSS.
 - b. Instead, the number of theta values within (-10, 10) one decimal point is used to estimate the number of scale points needed. This method yields 200 score units.

There are four achievement levels defined for the Maine scale: *Well-Below State Expectations*, *Below State Expectations*, *At State Expectations*, and *Above State Expectations*. The two fixed cuts are set at the *At State Expectations* and *Above State Expectations* levels. Table 4.7 presents the scaling constants, scale score cuts, and LOSS/HOSS.

Table 4.7. Maine Grade-Level Scale Properties

Grade	Scaling Constants		Scale Score Cuts			Range	
	Intercept	Slope	<i>Below State Expectations</i>	<i>At State Expectations</i>	<i>Above State Expectations</i>	LOSS	HOSS
Reading							
3	1507.14	11.90	1483	1500	1525	1400	1600
4	1503.75	12.50	1486	1500	1525	1400	1600
5	1505.26	13.16	1487	1500	1525	1400	1600
6	1505.26	13.16	1486	1500	1525	1400	1600
7	1502.63	13.16	1483	1500	1525	1400	1600
8	1500.00	12.50	1484	1500	1525	1400	1600
HS	1501.56	15.63	1489	1500	1525	1400	1600
Mathematics							
3	1511.25	12.50	1486	1500	1525	1400	1600
4	1507.00	10.00	1488	1500	1525	1400	1600
5	1502.08	10.42	1484	1500	1525	1400	1600
6	1501.14	11.36	1481	1500	1525	1400	1600
7	1505.44	10.87	1482	1500	1525	1400	1600
8	1504.35	10.87	1484	1500	1525	1400	1600
HS	1523.22	17.86	1489	1500	1525	1400	1600

Section 5: Technical Quality-Validity

Validity is defined by the *Standards for Educational and Psychological Testing* as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests” (AERA et al., 2014, p. 11). Validating a test score interpretation is not a quantifiable property but an ongoing process, beginning at initial conceptualization of the construct and continuing throughout the entire process of assessment development and implementation. Every aspect of an assessment development and administration provides evidence in support of (or a challenge to) the validity of the intended inferences about what students know based on their score, including design, content specifications, item development, test constraints, psychometric quality, standard setting, and administration.

As this technical report has progressed, it has covered the different phases of the testing cycle and provided different pieces of technical quality evidence. It provides relevant evidence and a rationale in support of test score interpretations and intended uses based on the *Standards*, which are considered to be “the most authoritative statement of professional consensus regarding the development and evaluation of educational and psychological tests” (Linn, 2006, p.27). The validity argument begins with a statement of the assessment’s intended purposes, followed by the evidentiary framework, where available validity evidence is provided to support the argument that the test actually measures what it purports to measure (SBAC, 2016).

First, the Through Year Assessment went through psychometric analyses—such as test reliability, classification accuracy, conditional standard error of measurement (CSEM), test information, differential item function (DIF), and convergent validity check—and the results so far strongly support the reliability and validity claims of this assessment. In addition, the test-development process ensures validity of the intended test score interpretations provided through the scale score. Last but not least, this assessment is aligned to grade-level content, and test scores are suitable for use in accountability systems as a result of a robust development process to determine the test blueprint, passage and item specifications, and ALDs.

5.1. Validity Evidence Framework

The *Standards* describes validation as a process of constructing and evaluating arguments for the intended interpretation and use of test scores:

“A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. . .

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system” (AERA et al., 2014, pp. 21–22).

The *Standards* (AERA et al., 2014, pp. 13–19) outlines the following five main sources of validity evidence:

- Evidence based on test content
- Evidence based on response processes

- Evidence based on internal structure
- Evidence based on relations to other variables
- Evidence based on validity and consequences of testing

Evidence based on test design refers to traditional forms of content validity or content-related evidence. Evidence based on response processes refers to the cognitive process engaged in by students when answering test items, or the “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers” (AERA et al., 2014, p. 15). Evidence based on internal structure refers to the psychometric analyses of “the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (AERA et al., 2014, p. 16). Evidence based on relations to other variables refers to traditional forms of criterion-related validity evidence, such as predictive and concurrent validity. Evidence based on validity and consequences of testing refers to the evaluation of the intended and unintended consequences associated with a testing program.

Table 5.1 presents an overview of the validity components covered in this technical report.

Table 5.1. Sources of Validity Evidence for Each Test Purpose

Test Purpose	Sources of Validity Evidence		
	Test Content	Response Processes	Internal Structure
1. To report individual student achievement relative to the state-adopted content standards in reading and mathematics	✓	✓	✓
2. To provide information to the public about school performance through the state’s Every Student Succeeds Act (ESSA) reporting system, the ESSA Dashboard	✓	✓	✓
3. To support school identification within the state’s ESSA compliant system of school identification and support	✓	✓	✓
4. To provide a source of information for ongoing local program evaluation	✓	✓	✓

5.2. Purposes and Evidence

5.2.1. Test Purpose 1

Purpose: To report individual student achievement relative to the state-adopted content standards in reading and mathematics

Sources of Validity Evidence Based on Test Content:

- Test blueprint, content specifications, and item specifications are aligned to the full breadth and depth of grade-level content, process skills, and associated cognitive complexity.
- Blueprint specifications are evaluated for each test event for regular and accommodated populations. The evaluations are performed prior to test administration by simulation and then again following test administration.

- For high school, tests are linked to the Maine Learning Results by the incorporation of the CCSS into item- and test-development specifications.
- Bias is minimized through Universal Design and accessibility resources.
- The item pool and item-selection procedures adequately support the test design.
- Operational computer adaptive test events meet all blueprint constraints, both for the general student population and for students taking accommodated test forms.
- Relevant sections within this report: 2, 3, 7, 8

Sources of Validity Evidence Based on Response Processes:

- Item-development and quality-control processes include screening and reviewing field test items for potential construct-irrelevant difficulty due to bias against demographic groups.
- The item types used in the assessment require response processes specified in the CCSS.
- The standard setting process relies on stakeholder judgments about proficiency based on student responses to, and the response processes elicited by, test items.
- Relevant sections within this report: 2, 7, 8

Sources of Validity Evidence Based on Internal Structure:

- ALDs were developed in consultation with committees of Maine educators with a secondary goal of providing information to all Maine educators.
- Achievement levels were set consistent with best practices through the embedded standard setting procedures.
- The assessment supports precise measurement and consistent classification to support analysis and reporting of scores.
- Item-calibration results support good fit of the test model and intended internal structure of the measurement construct.
- Differential Item Functioning (DIF) analysis was completed for all items across identifiable subgroups of students.
- Tests reliably measure on a scale that is established by achievement levels at every grade and reliably classify students into the achievement levels.
- Relevant sections within this report: 3, 4, 6, 8

5.2.2. Test Purpose 2

Purpose: To provide information to the public about school performance through the state's Every Student Succeeds Act (ESSA) reporting system, the ESSA Dashboard

Sources of Validity Evidence Based on Test Content:

- Test content is aligned with the reporting requirements of Maine's ESSA Dashboard.
- Bias is minimized through Universal Design and accessibility resources.
- Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity.
- The item pool and item-selection procedures adequately support the test design.
- Reporting categories align with the structure of Maine's standards to support the interpretation of the test results.
- Relevant sections within this report: 2, 8

Sources of Validity Evidence Based on Response Process:

- Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity.
- Achievement levels were set consistent with best practices.
- Relevant sections within this report: 2, 4, 7

Sources of Validity Evidence Based on Internal Structure:

- The assessment supports precise measurement and consistent classification to support analysis and reporting of longitudinal data.
- Reporting categories align with the structure of Maine's standards to support the interpretation of test results.
- Achievement levels were set consistent with best practices.
- Item-calibration results support good fit of the test model and intended internal structure of the measurement construct.
- Differential Item Functioning (DIF) analysis was completed for all items across identifiable subgroups of students.
- Relevant sections within this report: 2, 3, 4, 6

5.2.3. Test Purpose 3

Purpose: To support school identification within the state's ESSA compliant system of school identification and support

Sources of Validity Evidence Based on Test Content:

- Maine's model of school support emphasizes the importance of measurement for academic achievement and progress of English language arts and math.
- Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity.
- Reporting categories align with the structure of Maine's standards to support the interpretation of the test results.
- Relevant sections within this report: 2

Sources of Validity Evidence Based on Response Process:

- Bias is minimized through Universal Design and accessibility resources.
- Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity.
- Achievement levels are vertically articulated.
- Relevant sections within this report: 2, 3, 4, 6

Sources of Validity Evidence Based on Internal Structure:

- The assessment supports precise measurement and consistent classification to support analysis and reporting of longitudinal data.
- Achievement levels are vertically articulated.
- Item-calibration results support good fit of the test model and intended internal structure of the measurement construct.
- Differential Item Functioning (DIF) analysis was completed for all items across identifiable subgroups of students.
- Relevant sections within this report: 3, 4, 6

5.2.4. Test Purpose 4

Purpose: To provide a source of information for ongoing local program evaluation

Sources of Validity Evidence Based on Test Content:

- Reporting categories align with the structure of Maine’s standards to support the interpretation of the test results.
- Relevant sections within this report: 2, 8

Sources of Validity Evidence Based on Response Process:

- Bias is minimized through Universal Design and accessibility resources.
- Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity.
- ALDs were developed in consultation with committees of Maine educators with a secondary goal of providing information to all Maine educators.
- Relevant sections within this report: 2, 3, 4, 6, 8

Sources of Validity Evidence Based on Internal Structure:

- The assessment supports precise measurement and consistent classification for all students.
- ALDs were developed in consultation with committees of Maine educators with a secondary goal of providing information to all Maine educators.
- Scale is vertically articulated and supports longitudinal tracking of students’ academic progress.
- Achievement levels are vertically articulated.
- Relevant sections within this report: 2, 3, 4, 6

5.3. Interpretive Argument Claims

The test scores for the spring administration support their intended purposes. Claims to support this are documented in the technical report, as shown in Table 5.2.

Table 5.2. Interpretive Argument Claims, Evidence to Support the Essential Validity Elements

Argument	Tech Report Section(s)	Evidence
Tests and items were carefully developed to ensure that the test measured the Maine content standards.	2. Test Design and Development	Description of the development and review process for items, passages, and tests
Test score interpretations are comparable across students.	3.3. Constraint-Based Adaptive Test Engine 4. Item Statistics, Calibration, and Scaling 6. Technical Quality-Other	Simulations, analysis of test information, conditional standard errors of measurement, classification accuracy, and reliability estimates; blueprint comparability across students; item analysis, calibration, and scaling procedures
Test administrations were secure and standardized.	3. Administration and Security	Test administration procedures, including administration training, test accommodations, test security, and availability of help desk during testing window

Argument	Tech Report Section(s)	Evidence
Scoring was standardized and accurate.	6.4. Scoring 8.3. Reporting	Scoring rules and procedures; quality control of operational scoring
Achievement standards were rigorous and technically sound.	8. Achievement Standards and Reporting	Documentation of standard-setting procedures, including the methodology, identification of workshop participants, implementation process, and ALD development and validation
Assessments were accessible to all students and fair across student subgroups.	2. Test Design and Development 3. Administration and Security 6. Technical Quality-Other 7. Inclusion of all Students	Accommodation policy and implementation, sensitivity review, availability of translations, and DIF analyses

5.4. Validity Argument

The test development and technical quality of the Maine Through Year Assessment supports the intended test score interpretations that are provided through the scale scores and ALDs. The test blueprints, passage specifications, item specifications, and ALD development process show that the Maine Through Year Assessment is aligned to grade-level content standards. As an added dimension for adaptive testing, this assessment demonstrated that the tests administered to students conformed to the blueprint during the CBE evaluation studies.

The item pool and item-selection procedures used for the adaptive administration adequately support the test design and blueprint. Content experts developed expanded item types that allow response processes to reveal skills and knowledge. All items were carefully reviewed through multiple cycles of the item-development process for ambiguity, bias, sensitivity, irrelevant clues, and inaccuracy to ensure the fit between the construct and the nature of performance.

Studies for evidence based on relations to other variables and evidence based on consequences of testing have not been included within the scope of work undertaken to date by NWEA. This evidence may be added in future studies, such as evaluation of the concurrent validity of the assessment with external measures, evaluation of the effects of testing on instruction, evaluation of the effects of testing on issues such as high school dropout rates, analyses of students’ opportunity to learn, and analyses of changes in textbooks and instructional approaches (SBAC, 2016). The evaluation of unintended consequences may include changes in instruction, diminished morale among teachers and students, increased pressure on students leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging (SBAC, 2016).

Teacher surveys or focus groups can be used to collect information regarding the use of the tests and how the tests impacted the curriculum and instruction. A better understanding of the extent to which performance gains on assessments reflect improved instruction and student learning (rather than more superficial interventions such as narrow test-preparation activities) would also provide evidence based on consequences of test use. Longitudinal test data, along with additional information collected from educators (e.g., information on understanding of learning standards, motivation and effort to adapt the curriculum and instruction to content

standards, instructional practices, classroom assessment format and content, use and nature of test assessment preparation activities, and professional development), would allow for meaningful analyses and interpretations of the score gain and uniformity of standards, learning expectations, and consequences for all students.

Section 6: Technical Quality-Other

The *Standards for Educational and Psychological Testing* refers to reliability as the “consistency of scores across replications of a testing procedure” (AERA et al., 2014, p. 33). The level of reliability/precision of scores has implications for validity. In other words, scores must be consistent and precise enough to be useful for their intended purposes. If scores are to be meaningful, tests should produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory of the test. In addition, the range of certainty around the scores should be small enough to support educational decisions. The reliability/precision of the assessment was examined through analyses of measurement error under simulated and operational conditions, as follows:

- Marginal reliability for adaptive tests
- Cronbach’s alpha and standard error of measurement (SEM) for fixed forms
- Classification accuracy

Combined, these data provide several ways of looking at the reliability of student scores on a test. Classification accuracy provides important information related to achievement level classifications. These are of particular interest in the context of state accountability requirements.

6.1. Reliability

6.1.1. Marginal Reliability for Adaptive Tests

Traditional reliability coefficients from classical test theory consider individual items and depend on all test takers to take common items; however, in a CAT, different students receive different items. Therefore, the marginal reliability coefficient for the CAT administration was calculated. Samajima (1994) recommends the marginal reliability coefficient because it uses test information (e.g., variance of estimated theta and SEM) to estimate the reliability of student scores:

$$\text{Marginal Reliability} = \frac{\text{var}(\hat{\theta}) - \sigma^2}{\text{var}(\hat{\theta})}$$

where σ is defined as:

$$\sigma = E\{[I(\theta)]^{-1/2}\}$$

Table 6.1 and Table 6.2 present the overall error of estimated theta and test reliability for the grades 3–8 adaptive tests. Each table includes the average number of items administered, the standard deviation (SD) of the estimated theta, the mean conditional standard error of measurement (CSEM), and the marginal reliability coefficient. The SD of estimated theta and mean SEM are relatively small, and the marginal reliability of the overall scores is 0.93 or higher for reading and 0.95 or higher for math. These results indicate that, overall, the score precision is reasonable: the overall mean SEM values were approximately 0.40, while the reliability estimates are consistent with the guidelines for reliability in a graduation test (Phillips & Camara, 2006).

Table 6.1. Reliability Statistics—Reading

Grade	Average # Items	SD of Estimated Theta	Mean SEM	Reliability
3	27	1.46	0.41	0.95
4	27	1.30	0.40	0.94
5	27	1.28	0.41	0.93
6	27	1.22	0.41	0.93
7	27	1.33	0.40	0.94
8	27	1.29	0.41	0.93

Table 6.2. Reliability Statistics—Mathematics

Grade	Average # Items	SD of Estimated Theta	Mean SEM	Reliability
3	27	1.65	0.41	0.96
4	27	1.77	0.40	0.97
5	27	1.69	0.40	0.97
6	27	1.72	0.40	0.97
7	27	1.71	0.40	0.97
8	27	1.46	0.41	0.95

6.1.2. Reliability for HS Fixed Forms

Cronbach's alpha reliability coefficient is a frequently used measure of internal consistency over the responses to a set of items measuring an underlying, unidimensional trait. Reliability coefficient alpha expresses the consistency of test scores as the ratio of true score variance to total score (observed) variance (true score variance + error variance). A larger index would indicate that test scores were less influenced by random sources of error. The reliability coefficient is a "unitless" index, which can be compared from test to test and ranges from 0.0 to 1.0, where 0.80 is typically considered the minimally acceptable level of reliability for assessments. While sensitive to random error associated with content sampling variability, the index is not sensitive to other types of errors, such as temporal stability or variability in performance that might occur across different testing occasions. Cronbach's alpha is computed as follows (Crocker & Algina, 1986):

$$\hat{\alpha} = \frac{k}{k-1} \left(1 - \frac{\sum \sigma_j^2}{\sigma_x^2} \right)$$

where k is the number of items, σ_x^2 is the total score variance, and σ_j^2 is the variance of item j .

The SEM is an index of the random variability in test scores in raw score units and is defined as:

$$\text{SEM} = SD\sqrt{1 - \hat{\alpha}}$$

where SD represents the standard deviation of the raw score distribution, and $\hat{\alpha}$ represents Cronbach's alpha. The overall SEM is expressed in raw score units and is a test-level statistic. Table 6.3 presents Cronbach's alpha reliability coefficients, along with the SEMs.

Table 6.3. Cronbach's Alpha (Internal Consistency) for Fixed Forms

Content Area	#Items	Reliability	SEM
Reading	30	0.86	4.98
Mathematics	30	0.80	5.15

6.1.3. Classification Accuracy

Classification accuracy is a measure of how accurately test scores place students into reporting category levels. It refers to the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores. It is common to estimate classification accuracy by using a psychometric model to find true scores corresponding to observed scores. The likelihood of inaccurate placement depends on the amount of error associated with scores, especially those nearest cut points.

Classification accuracy was calculated as follows (SBAC, 2016):

1. For each student, a normal distribution was constructed, with means equal to the scale score estimate and standard deviation equal to the SEM as a plausible true score distribution.
2. For each student, the proportion of that normal distribution that fell within each achievement level was calculated.
3. Within the groups of students assigned to a particular achievement level (Level 4, 3, 2, or 1 for the overall score), the sums of the proportions over students were computed. This provided estimates of the number of students whose true score falls within a level for each assigned achievement level. These sums were then expressed as a proportion of the total sample (i.e., expected proportion).
4. With the table of expected proportions, correct classification rates were then defined. This is the proportion of students whose true classification agrees with the assigned level among the subset of students with that assigned level.
5. The overall classification rate is the sum of the proportions of students whose true score level agrees with the assigned level divided by the total proportion of students assigned to a level.

Table 6.4 and Table 6.5 present the classification accuracy results by grade and achievement level. Overall classification accuracy for adaptive tests ranges from 0.80 (grade 4 reading) to 0.86 (grade 7 mathematics), whereas the values for high school fixed-form tests are 0.78 and 0.66 for reading and mathematics, respectively. With the plan being for high school to be adaptive in the coming years, the classification accuracy results are expected to be higher than this year's. In general, classification accuracy is moderate to high.

Table 6.4. Classification Accuracy by Achievement Level—Reading

Grade	Achievement Level	N	Prop.	Expected Proportion ^a				Class. Acc.	Overall Class. Acc.
				L1	L2	L3	L4		
3	<i>Well-Below State Expectations</i>	1531	0.13	0.11	0.02	0.00	0.00	0.85	0.81
	<i>Below State Expectations</i>	3272	0.27	0.03	0.21	0.03	0.00	0.78	
	<i>At State Expectations</i>	5723	0.47	0.00	0.05	0.39	0.03	0.83	
	<i>Above State Expectations</i>	1565	0.13	0.00	0.00	0.03	0.10	0.77	
4	<i>Well-Below State Expectations</i>	1467	0.12	0.10	0.02	0.00	0.00	0.83	0.80
	<i>Below State Expectations</i>	2881	0.24	0.03	0.17	0.04	0.00	0.71	
	<i>At State Expectations</i>	5854	0.49	0.00	0.04	0.41	0.03	0.84	
	<i>Above State Expectations</i>	1858	0.15	0.00	0.00	0.03	0.12	0.80	
5	<i>Well-Below State Expectations</i>	1519	0.13	0.11	0.02	0.00	0.00	0.85	0.81
	<i>Below State Expectations</i>	2198	0.19	0.03	0.13	0.03	0.00	0.68	
	<i>At State Expectations</i>	6283	0.53	0.00	0.04	0.45	0.03	0.85	
	<i>Above State Expectations</i>	1853	0.16	0.00	0.00	0.03	0.12	0.75	
6	<i>Well-Below State Expectations</i>	1258	0.10	0.09	0.01	0.00	0.00	0.90	0.81
	<i>Below State Expectations</i>	2710	0.23	0.03	0.16	0.03	0.00	0.70	
	<i>At State Expectations</i>	6440	0.53	0.00	0.06	0.45	0.03	0.85	
	<i>Above State Expectations</i>	1633	0.14	0.00	0.00	0.02	0.11	0.79	
7	<i>Well-Below State Expectations</i>	1392	0.11	0.10	0.02	0.00	0.00	0.91	0.82
	<i>Below State Expectations</i>	3035	0.25	0.03	0.19	0.03	0.00	0.76	
	<i>At State Expectations</i>	6139	0.50	0.00	0.05	0.42	0.03	0.84	
	<i>Above State Expectations</i>	1622	0.13	0.00	0.00	0.02	0.11	0.85	
8	<i>Well-Below State Expectations</i>	1279	0.10	0.09	0.02	0.00	0.00	0.90	0.82
	<i>Below State Expectations</i>	3045	0.24	0.03	0.18	0.03	0.00	0.75	
	<i>At State Expectations</i>	6710	0.53	0.00	0.05	0.45	0.03	0.85	
	<i>Above State Expectations</i>	1547	0.12	0.00	0.00	0.03	0.10	0.83	
HS	<i>Well-Below State Expectations</i>	1624	0.13	0.11	0.02	0.00	0.00	0.85	0.78
	<i>Below State Expectations</i>	3007	0.25	0.05	0.16	0.04	0.00	0.64	
	<i>At State Expectations</i>	6037	0.50	0.00	0.05	0.42	0.03	0.84	

Grade	Achievement Level	N	Prop.	Expected Proportion ^a				Class. Acc.	Overall Class. Acc.
				L1	L2	L3	L4		
	<i>Above State Expectations</i>	1490	0.12	0.00	0.00	0.03	0.09	0.75	

^a Level 1 = *Well-Below State Expectations*, Level 2 = *Below State Expectations*, Level 3 = *At State Expectations*, and Level 4 = *Above State Expectations*.

Table 6.5. Classification Accuracy by Achievement Level—Mathematics

Grade	Achievement Level	N	Prop.	Expected Proportion ^a				Class. Acc.	Overall Class. Acc.
				L1	L2	L3	L4		
3	<i>Well-Below State Expectations</i>	2100	0.17	0.15	0.02	0.00	0.00	0.88	0.82
	<i>Below State Expectations</i>	2562	0.21	0.03	0.15	0.03	0.00	0.71	
	<i>At State Expectations</i>	5333	0.44	0.00	0.05	0.37	0.02	0.84	
	<i>Above State Expectations</i>	2156	0.18	0.00	0.00	0.02	0.15	0.83	
4	<i>Well-Below State Expectations</i>	2252	0.19	0.17	0.02	0.00	0.00	0.89	0.84
	<i>Below State Expectations</i>	2979	0.25	0.03	0.18	0.03	0.00	0.72	
	<i>At State Expectations</i>	5346	0.44	0.00	0.04	0.38	0.02	0.86	
	<i>Above State Expectations</i>	1561	0.13	0.00	0.00	0.02	0.11	0.85	
5	<i>Well-Below State Expectations</i>	2206	0.19	0.16	0.02	0.00	0.00	0.84	0.84
	<i>Below State Expectations</i>	3651	0.31	0.03	0.25	0.03	0.00	0.81	
	<i>At State Expectations</i>	4770	0.40	0.00	0.04	0.34	0.02	0.85	
	<i>Above State Expectations</i>	1292	0.11	0.00	0.00	0.02	0.09	0.82	
6	<i>Well-Below State Expectations</i>	2269	0.19	0.17	0.02	0.00	0.00	0.89	0.84
	<i>Below State Expectations</i>	4397	0.36	0.04	0.29	0.04	0.00	0.81	
	<i>At State Expectations</i>	4341	0.36	0.00	0.05	0.30	0.01	0.83	
	<i>Above State Expectations</i>	1073	0.09	0.00	0.00	0.01	0.08	0.89	
7	<i>Well-Below State Expectations</i>	2463	0.20	0.19	0.01	0.00	0.00	0.95	0.86
	<i>Below State Expectations</i>	4408	0.36	0.03	0.30	0.03	0.00	0.83	
	<i>At State Expectations</i>	4338	0.35	0.00	0.04	0.30	0.01	0.86	
	<i>Above State Expectations</i>	1044	0.09	0.00	0.00	0.01	0.07	0.78	
8	<i>Well-Below State Expectations</i>	2590	0.21	0.18	0.02	0.00	0.00	0.86	0.83
	<i>Below State Expectations</i>	4939	0.39	0.05	0.31	0.04	0.00	0.79	

Grade	Achievement Level	N	Prop.	Expected Proportion ^a				Class. Acc.	Overall Class. Acc.
				L1	L2	L3	L4		
	<i>At State Expectations</i>	4224	0.33	0.00	0.04	0.28	0.01	0.85	
	<i>Above State Expectations</i>	872	0.07	0.00	0.00	0.01	0.06	0.86	
HS	<i>Well-Below State Expectations</i>	3082	0.25	0.18	0.06	0.01	0.00	0.72	0.66
	<i>Below State Expectations</i>	3949	0.32	0.08	0.17	0.08	0.00	0.53	
	<i>At State Expectations</i>	4372	0.36	0.01	0.07	0.26	0.02	0.72	
	<i>Above State Expectations</i>	789	0.06	0.00	0.00	0.01	0.05	0.83	

^a Level 1 = *Well-Below State Expectations*, Level 2 = *Below State Expectations*, Level 3 = *At State Expectations*, and Level 4 = *Above State Expectations*.

6.2. Fairness and Accessibility

Assessment fairness and accessibility are addressed in multiple approaches in this report. First, Universal Design is used to design the test and items (see Section 2.3.1). Second, accommodations are provided according to special student needs during administration and through various paper forms (Section 3.4 and Section 7). Third, analyses are conducted to evaluate item fairness and accessibility. While the first two approaches are qualitative methods, the last approach is quantitative. This section addresses the methods and results of these analyses.

Differential item functioning (DIF) is a statistical procedure that flags items for potential bias. The fundamental measurement assumption of DIF is that the probability of a correct response to a test item is a function of the item's difficulty and the student's ability. This function is expected to remain invariant to other characteristics unrelated to ability, such as gender and ethnicity. Therefore, if two students with the same ability respond to the same item, they are assumed to have an equal probability of answering the item correctly. To test this assumption, responses to items by students sharing an aspect of a characteristic (e.g., gender) are compared with responses to the same items by other students who share a different aspect of the same characteristic (e.g., males vs. females). The group representing students in a specific demographic group is referred to as the *focal* group. The group comprised of students from outside this group is referred to as the *reference* group.

When DIF is detected and the fundamental measurement assumption does not hold (i.e., students with the same ability in different groups of interest have different probabilities of correctly answering an item), the item is said to be functioning differently for the two groups. The presence of DIF in an item suggests that the item is functioning unexpectedly regarding the groups included in the comparison. The cause of the unexpected functioning is not revealed in a DIF analysis. It may be that item content is inadvertently providing an advantage or disadvantage to members of one of the two groups. Content experts who have special knowledge of the groups involved can often identify a cause of this type. DIF may also result from differential instruction closely associated with group membership.

Because fairness is a fundamental validity issue, it is essential that items be reviewed and assessed for DIF. Many methods for assessing DIF have been used and compared in conventional paper-pencil tests; however, DIF detection may be more important for a CAT than it is for traditional paper-pencil tests for two reasons (Zwick et al., 1994): First, items with DIF may be more consequential for the examinees because fewer items are administered in a CAT. Second, several potential sources of DIF may be introduced, such as differential computer familiarity, facility, and anxiety. The difficulty of DIF analysis in a CAT is introduced by the fact that different sets of items are administered to different examinees. Therefore, the logistic regression (LR) procedure was applied to items that were administered in this CAT. The LR is also used on the HS summative items, although they are on a fixed form.

6.2.1. Logistic Regression (LR) DIF Method

The LR DIF procedure models item responses (for both dichotomous and polytomous items) as a function of group memberships, ability estimates, and their interaction. Testing for the presence of DIF based on logistic regression provide a model-based approach to identify uniform and non-uniform DIF. DIF is classified as uniform if the effect is constant; that is, uniform DIF exists when the difference in the probabilities of a correct answer for the two groups is the same at all ability levels. DIF is classified as non-uniform if the effect varies conditional on

the ability level; that is, non-uniform DIF exists if the interaction between item-response function and group membership is disordinal.

The LR procedure compares the following three models (Fu & Monfils, 2016; Swaminathan & Rogers, 1990; Zumbo, 1999):

$$\begin{aligned} \text{Model 1: } \text{logit}(P) &= \beta_0 + \beta_1 X + \beta_2 E \\ \text{Model 2: } \text{logit}(P) &= \beta_0 + \beta_1 X + \beta_2 G + \beta_3 E \\ \text{Model 3: } \text{logit}(P) &= \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG + \beta_4 E \end{aligned}$$

where:

- P is the probability of a test taker answering an item incorrectly (for a dichotomous item) and the probability of getting an item score or lower (for a polytomous item).
- X is the criterion variable (typically an ability estimate).
- G is the group membership.
- E is a vector, including additional explanatory variables.
- β are the associated regression parameters for model k .

For both dichotomous and polytomous items, Models 1, 2, and 3 are also referred to as a no-DIF model, a uniform DIF model, and a nonuniform DIF model, respectively. The group estimates (β_2) are related to uniform DIF, and the interaction estimates (β_3) are associated with nonuniform DIF. Note that for a dichotomously scored item, the target probability that the LR estimates is the probability of answering an item incorrectly, which is different from the probability of answering an item correctly that many people may be accustomed to. Similarly, the target probability in the regression model for a polytomously scored item is the probability of obtaining an item score or below, to be consistent with that for a dichotomously scored item.

The item shows DIF if the modeled fit statistic is improved when group and interaction are added to the model, in order. To test the presence of nonuniform DIF, Model 2 and Model 3 are compared, using the likelihood ratio test with 1 degree of freedom (df) in chi-square distribution:

$$x^2 = [-2 \ln L(\text{Model2})] - [-2 \ln L(\text{Model3})]$$

Similarly, to test the presence of uniform DIF, Model 1 and Model 2 are compared, using the likelihood ratio test with 1 df:

$$x^2 = [-2 \ln L(\text{Model1})] - [-2 \ln L(\text{Model2})]$$

To test overall DIF (uniform DIF or nonuniform DIF), Model 1 and Model 3 are compared, using the likelihood ratio test with 2 df:

$$x^2 = [-2 \ln L(\text{Model1})] - [-2 \ln L(\text{Model3})]$$

The effect size is also used to avoid practically trivial but statistically significant results (French & Miller, 1996). Effect size is indicated by the difference of the Nagelkerke R^2 between two models (Gómez-Benito et al., 2009). Table 6.6 presents the DIF classification rules for the LR DIF procedure. These rules were confirmed to be consistent to the Mantel-Haenszel DIF classification rule for dichotomous items used by ETS (Fu & Monfils, 2016).

Table 6.6. LR DIF Categories

DIF Category	Level of DIF	Definition
A	Negligible	χ^2 test is not significant at 0.05 level or $\Delta R^2 < 0.035$.
B	Moderate	χ^2 test is significant at 0.05 level and $0.035 \leq \Delta R^2 < 0.070$.
C	Strong	χ^2 test is significant at 0.05 level and $\Delta R^2 \geq 0.070$.

Note. ΔR^2 is the Nagelkerke R^2 difference between two models.

6.2.2. DIF Results

DIF analysis is performed between a pair of demographic subgroups, typically defined by gender or ethnicity. For gender, male was used for the reference group, and female was used for the focal group; for ethnicity, white was used for the reference group, and a different minority subgroup was used for the focal group. More than 80% of students are white for the spring test. The large discrepancy in counts between reference group and focal group may cause statistical bias in estimates. Therefore, DIF was not conducted if the sample size for either group was less than 100. There are reduced counts of adaptive items meeting the minimum student counts required for DIF analyses due to the nature of adaptive item selection, while field test items were controlled to have required student counts and to be distributed across demographic groups.

Table 6.7 and Table 6.8 present the number of items identified for DIF for operational items and field test items, respectively. Considering that the Rasch model is applied (i.e., the same slope is assumed for all items), uniform DIF results are reported. The “+” sign next to the DIF category indicates that the item is in favor of the reference group, and the “-” sign indicates that the item is in favor of the focal group. As shown in the tables, most items were classified into Category A DIF, indicating negligible differential item functioning. Among the items eligible for DIF screening, the maximum proportion of items displaying Category B DIF did not exceed 1.5% per grade. Typically, item review is focused on items classified as exhibiting Category C DIF; no such DIF items were found in the item pool.

Table 6.7. DIF Analysis Results—Operational Items

Grade	Focal Group	Item Count by DIF Category					
		Total	A	B+	B-	C+	C-
Reading							
3	Female	304	303	–	1	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	33	33	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	26	26	–	–	–	–
4	Female	320	320	–	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	25	25	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	20	20	–	–	–	–
5	Female	372	369	3	–	–	–
	Black or African American	2	2	–	–	–	–

Grade	Focal Group	Item Count by DIF Category					
		Total	A	B+	B-	C+	C-
	Hispanic	16	16	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	17	17	-	-	-	-
6	Female	287	287	-	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	35	35	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	21	21	-	-	-	-
7	Female	286	285	1	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	25	25	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	13	13	-	-	-	-
8	Female	321	320	1	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	25	25	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	13	13	-	-	-	-
HS	Female	30	30	-	-	-	-
	Black or African American	30	30	-	-	-	-
	Hispanic	30	30	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	30	30	-	-	-	-
Mathematics							
3	Female	457	456	1	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	23	23	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	9	9	-	-	-	-
4	Female	326	326	-	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	17	17	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	3	3	-	-	-	-
5	Female	400	397	2	1	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	28	28	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	4	4	-	-	-	-
6	Female	342	337	5	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	28	28	-	-	-	-

Grade	Focal Group	Item Count by DIF Category					
		Total	A	B+	B-	C+	C-
	Asian	–	–	–	–	–	–
	Two or More Races	3	3	–	–	–	–
7	Female	326	325	1	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	42	42	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	11	11	–	–	–	–
8	Female	350	349	1	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	38	38	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	9	9	–	–	–	–
HS	Female	30	30	–	–	–	–
	Black or African American	30	30	–	–	–	–
	Hispanic	30	30	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	30	30	–	–	–	–

Table 6.8. DIF Analysis Results—Field Test Items

Grade	Focal Group	Item Count by DIF Category					
		Total	A	B+	B-	C+	C-
Reading							
3	Female	60	60	–	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	7	7	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	7	7	–	–	–	–
4	Female	126	126	–	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
5	Female	98	98	–	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
6	Female	87	87	–	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	17	17	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	6	6	–	–	–	–

Grade	Focal Group	Item Count by DIF Category					
		Total	A	B+	B-	C+	C-
7	Female	137	137	-	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	-	-	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-
8	Female	130	130	-	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	-	-	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-
HS	Female	19	19	-	-	-	-
	Black or African American	5	5	-	-	-	-
	Hispanic	7	7	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	5	5	-	-	-	-
Mathematics							
3	Female	160	160	-	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	-	-	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-
4	Female	153	153	-	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	-	-	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-
5	Female	160	160	-	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	-	-	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-
6	Female	160	159	-	1	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	-	-	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-
7	Female	159	159	-	-	-	-
	Black or African American	-	-	-	-	-	-
	Hispanic	-	-	-	-	-	-
	Asian	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-
8	Female	159	157	2	-	-	-

Grade	Focal Group	Item Count by DIF Category					
		Total	A	B+	B-	C+	C-
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
HS	Female	37	37	–	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	12	12	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	2	2	–	–	–	–

6.3. Full Achievement Continuum

It is important for an assessment to cover the full achievement continuum in order to provide reliable scores of the entire score range, or at least at the cut scores to provide higher classification accuracy. The summative item bank covers a wide range of difficulties, as shown in Table 4.5. This enables the summative assessment to effectively differentiate between lower- and higher-performing students. Most importantly, it increases accuracy in classifying students’ achievement levels, especially for students just above or below the cut scores. The evidence on CSEMs from Section 3.3.3 indicates the tests can accurately estimate ability across the full ability scale, especially at the middle range of the scale and around the cut scores.

6.4. Scoring

There are two scoring approaches to estimate student scores: number correct and pattern scoring. The number correct method uses student responses to determine student scores: correct vs. incorrect for dichotomous items and earned score points for polytomous items. This method yields a one-to-one correspondence between raw scores and scale scores. Pattern scoring not only considers student responses but also item difficulty in score decisions. Answering a difficult item correctly will yield a higher score than answering an easier item correctly; thus, when two students earn the same raw scores through different item sets, their scale scores may differ because of the difference in difficulty between the two sets of items. Consequently, pattern scoring yields multiple correspondences between raw scores and scale scores.

The goal of computer adaptive testing is to reach a desirable score precision across the student’s ability range. Student ability estimates (thetas) are computed during test administration to select subsequent items that assist in obtaining reliable scores. Pattern scoring helps attain stable student ability estimates quicker than the number correct method because of the inclusion of item difficulty in estimation. Thus, it is typically used for an adaptive test.

6.4.1. Construct Maine Scale

Rationales and procedures for constructing the Maine Through Year Assessment are described in Section 4.4. Both literature and practical considerations play important parts in the procedures. The rationales and procedures were discussed with the TAC members. The TAC’s feedback was also considered when determining the scale properties. Achievement levels established on the Maine scale score are determined by the standard setting meeting and approved by the School Board (see Section 8).

6.4.2. Machine-Scored Items

The Maine Through Year Assessment has only machine-scored items. The item pool included technology-enhanced items and constructed-response items; however, those items typically have multiple correct answer keys. The keys have been evaluated, checked, and then hard coded into the database for scoring purposes. Calibration and validation of test item parameters were described in Section 4.2 and Section 4.3. Note that technology-enhanced items were excluded when constructing paper forms (including large print and braille forms) due to the limitations of the media.

6.4.3. Attemptedness Rule and Not-Tested Codes

Attemptedness for the Maine Through Year Assessment is defined as answering at least 25% of the summative items. With different test lengths across grades and content areas, a fixed value (7 items) is selected for all tests. Besides this attemptedness rule, there are also situations that could invalidate student test scores. Different Not-Tested Codes (NTC) are assigned to pinpoint different causes of score invalidation. Table 6.9 lists the various NTC codes. A student's Maine scale score and achievement level are not reported when the attemptedness threshold is not met or an NTC code is present.

Table 6.9. Available Not-Tested Codes

NTC Code	Description
INV	Invalid: The student's assessment was invalidated, such as due to a security breach or if the student refused to finish the test.
PAR	Parent Refusal: The student was not tested because of a written request from a parent or guardian.
STR	Student Refusal: The student was not tested due to the student's refusal to participate.
EMW	Emergency Medical Waiver: The student was not tested because of an approved emergency medical waiver.
RMV	Removal: The student left the district before the test window, the student is a full-time homeschooled student, or there are duplicate student records.

6.5. Multiple Assessment Forms

An adaptive test has a large item pool in comparison with the number of items used in a fixed-form test. Items administered to individual students are selected according to the students' responses to prior items. Each student may have received a different set of items at the completion of the test. In other words, an adaptive test has multiple test forms by nature.

In order for an adaptive test to work, an item bank with all items equated to a common scale is essential for selecting items according to student ability. All items used for the Maine Through Year Assessment were equated to a common scale prior to the Spring 2023 administration, using data that did not come from Maine students. Thus, post-equating is needed to build the Maine scale score using actual Maine students. Spring 2023 is the first administration of the Maine Through Year Assessment. Although large item pools are available for each grade (except HS), the item pool sizes were reduced to increase the number of students taking most of the items for calibration purposes. Items not used in the Spring 2023 administration and newly field test items will be used in future administrations.

6.6. Multiple Versions of an Assessment

The Maine Through Year Assessment is mostly an adaptive test, but various paper accommodation forms are built for students with special needs. The number of students taking paper forms is not large enough for calibration. Instead, item parameters are derived from the adaptive test. The parameters are then used to derive scores for students who took paper forms. This approach makes the scores of the adaptive test and paper forms comparable.

6.7. Technical Analysis and Ongoing Maintenance

When planning the Spring 2023 assessment, test blueprint, test design, item development, specifications for CBE setup, and various psychometric analyses were considered. The test design, procedures, and methods documented in this report were applied to the Spring 2023 administration and will continue be used as guidelines for maintaining test consistency across administrations.

Section 7: Inclusion of All Students

Multiple guides were created for the Maine Through Year Assessment to explain the target population, supports, and accommodations for all students or specific populations, as well as guidance for test coordination and administration. The guides provided include:

1. *The Maine Through Year Assessment Administration Checklist Spring 2023*
2. *The Maine Through Year Assessment Coordinator Guide*
3. *The Maine Through Year Assessment Administration Guide*
4. *The Maine Through Year Assessment Proctor User Guide*
5. *The Maine Through Year Assessment User and Student Management Guide*
6. *The Maine Through Year Assessment Accessibility Guide*
7. *NWEA State Solutions: NWEA System and Technology Guide*

7.1. Testing Population

The Maine Through Year Assessment Coordinator Guide states that the Maine Through Year Assessment is designed for students in grades 3–8 and their second year of high school, with the exception of students with the most significant cognitive disabilities who have been found eligible for alternate assessments via the IEP Team Process. It is expected that approximately 99% of the student population participates in the Maine Through Year Assessment. The Every Student Succeeds Act (ESSA, 2015) requires that at least 95% of students (who are eligible to test) participate in the state assessments.

Table 1.1 in Section 1.3 provides the number of students registered and the actual number of students who participated in the Spring 2023 Through Year Assessment by grade and content area.

7.2. Procedures for Including Students Who Utilize Accessibility Features

The Maine Through Year Assessment Coordinator Guide states that “All students are expected to participate in state assessments. No student, including students with disabilities, may be excluded from the state assessment and accountability system” (p. 9).

Three tiers of accessibility features have been developed to support the inclusion of all students, such as students with disabilities (SWDs): universal tools, designated supports, and accommodations (as described in Section 7.4).

7.3. Procedures for Including Multilingual Learners

In compliance with the Every Student Succeeds Act (ESSA, 2015) and state law on the inclusion of Multilingual Learners (MLs), *The Maine Through Year Assessment Coordinator Guide* states that “Districts should carefully consider the tools and resources utilized by MLs on a routine basis to access classroom instruction. These should be implemented as designated supports for the student during the assessment experience” (pp. 9 & 10).

Guidelines for the participation of newly arrived multilingual learners are also addressed in the guide.

7.4. Accommodations

Accommodations increase accessibility to a test by removing barriers without affecting the test construct. Accessibility is an important part of score validity, as student scores should represent the knowledge, skills, and abilities of the student. If a student cannot fully access the test, then the score cannot properly represent the individual's achievement. Accessibility to the test was considered at different stages of test development and administration.

At the development stage: Universal Design was used to guide item development and style (see Section 2.3.1 for more details). Content and Bias Review and Data Review meetings checked for potential item bias through qualitative and quantitative methods. In addition to the adaptive test, fixed-form standard print, large print, and braille forms were created for students with a documented need in an IEP or 504 Plan. During paper-based form creation, items were hand selected to ensure the blueprints were met at each grade level for each content area. Items were carefully sequenced and reviewed to avoid clueing within a grade level. The item types selected for the paper-based forms include multiple choice, multiselect, and composite (which uses elements of both multiple choice and multiselect). Additionally, items do not include any art that is inappropriate for the visually impaired population. As a back-up, the braille vendor will reach out to NWEA if something cannot be brailled, which did not happen this year. The psychometric team provided statistical targets to the content team and reviewed and approved all selections to ensure that items on the paper forms were of similar difficulty, complexity, and compatibility to those selected by the constraint-based engine for the adaptive tests.

At the administration stage: Universal tools were provided within the test platform and accessible by all students. Students have the choice to use any of the available tools. Some of the universal tools are embedded in the online secure browser and don't require activation, such as answer eliminator, zoom, guideline, calculator for select math items, etc. Scrap/scratch paper is a nonembedded universal tool required to be provided to all students by the proctor. Information on the use of the universal tools was not recorded.

Another tier of accessibility features is designated supports. Designated supports can be provided to students who meet the following two criteria:

1. An educational team with knowledge of the student's achievement has determined that the support is appropriate for the student.
2. The support is consistent with the student's routine instruction and assessment.

Text-to-Speech (TTS) is available as an embedded designated support that needs to be assigned within the assessment platform. Table 7.1 provides the number of students who used TTS. Other designated supports that cannot be embedded in the online system are made available by the test administrator/proctor, such as individual/separate setting, small group setting, alternate aids/supports, and bilingual word glossary. In addition to the paper-based form accommodations, other accommodations include read aloud, American sign language, scribe, calculator, and read aloud for passages.

Refer to *The Maine Through Year Assessment Accessibility Guide* for more details regarding universal tools, designated supports, and accommodations.

Table 7.1. Number of Students Who Used TTS

Grade	Content Area	Number of Students
3	Math	2,055
3	Reading	1,751
4	Math	1,997
4	Reading	1,692
5	Math	1,745
5	Reading	1,416
6	Math	1,546
6	Reading	1,348
7	Math	1,295
7	Reading	1,103
8	Math	1,233
8	Reading	1,088
10	Math	441
10	Reading	415




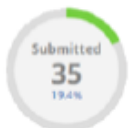
7.5. Monitoring Test Administration for Special Populations

Monitoring of the test administration is conducted in two ways: through the assessment administration and management system and through Maine DOE site visits.

7.5.1. Monitoring in Acacia

The Acacia system provides multiple pieces of information related to monitoring test status both during and after assessment. During the testing window, a testing status icon can be used to help proctors monitor student testing status with ease (Figure 7.1). After the testing window, the testing time marks at the item level can be analyzed to help understand the total test duration, time spent on each item, and any student test behavior related to testing time.

Figure 7.1. Monitoring Testing Status in Acacia

Icon	Assessment Status Icon Description
	<p>The Ready to Test icon displays the number and percentage of students who are enrolled and ready to take the assessment. It includes assessments in the Registered, Enrolled, and Ready to Test statuses. All assessments remaining in these statuses at the end of the assessment window are changed to Expired.</p>
 	<p>The In Progress icon displays the number and percentage of students actively testing. It includes assessments in the In Progress status only.</p> <p>The Alerts icon displays the number and percentage of students who have logged out and have not completed an assessment or have an enrollment hold. These students need test ticket login information to log back in and complete an assessment. This count includes assessments in the Inactive and Enrollment Hold statuses.</p> <p>Note: If any assessment registrations are in the Enrollment Hold status during the week before the assessment starts, contact NWEA Partner Support to resolve the hold.</p>
	<p>The Submitted icon displays the number and percentage of students who completed and submitted assessments. It includes assessments in the Submitted status only.</p>

7.5.2. Maine DOE Site Visits

In May 2023, during the assessment administration window, the Maine DOE Assessment Team conducted on-site visits at 14 School Administrative Units (SAUs) across the state of Maine. These on-site visits consisted of an observation of at least one assessment session in either reading and/or mathematics and a meeting with the on-site School Assessment Coordinator, District Assessment Coordinator, proctors, and/or other school personnel, as appropriate, to discuss pre-administration activities and planning, assessment security, accessibility features, proctor training, and SAU concerns or questions. On-site observations were completed using the Spring 2023 *Maine Through Year Assessment Observation Form* shown in Figure 7.2.

Figure 7.2. 2023 Maine Through Year Assessment Observation Form

Items indicated in ***bold italic font*** are areas of focus for the Spring 2023 Maine Through Year Assessment.

<i>School Name:</i>	
Assessment Administrator:	Proctor/TA/AA(s):
<i>Observer:</i>	<i>Subject:</i>
<i>Date of Observation:</i>	<i>Grade:</i>

	Item	Code*	Comments
1	Instructional materials that may provide clues or answers are not visible in the room.		
2	The desks/tables are arranged with enough space between them to minimize opportunities to review each other's work.		
3	Desks/tables are clear of all materials except what is allowed in the assessment administrator manual.		
4	Electronic devices were collected or otherwise stored away and unavailable for student use.		
5	The Assessment Administrator read directions clearly, loudly, and exactly as printed in the Assessment Administration Manual.		
6	Students worked independently of each other.		
7	The assessment room was free of disruptions (talking, fire drills, intercom announcements).		
8	Booklets/tickets were distributed to and collected from the students individually by the Assessment Administrator/Proctor(s) and not passed by students.		
9	The Assessment Administrator answered only questions related to the directions.		
10	Students were provided a break individually, (where applicable) during an assessment session with close supervision.		

11	Students worked on appropriate sections of the assessment and did not return to or go forward to other sections.		
12	All students remained quiet as everyone completed the assessment session.		
13	Assessment tickets/booklets, answer documents, and scrap paper were never left unattended.		
14	The assessment room was supervised at all times.		
15	The Assessment Administrator/Proctor(s) were actively monitoring the room at all times.		
16	Assessment signs were posted on room doors (e.g., Do Not Disturb, Electronic Devices Not Allowed, Quiet Please Assessments in Progress).		

* Use Codes: NA = Not Applicable; 1 = Exemplary; 2 = Acceptable; 3 = Minor Issue; 4 = Major Issue; UO = Unable to Observe

Is this the TA's first time administering the assessment?

- Yes
 No

TA's level of confidence administering the assessment.

- High
 Neutral
 Low

Does the proctor/TA/AA feel they received sufficient training and support to administer the assessment?

- Yes
 No

If no, please explain.

Did you observe any students or did the specifically observed student complete the entire assessment?

- Yes
 No

If no, please provide a reason why the student or students did not complete the assessment. Please check all that apply.

- Student became ill and left the room
 Student became overwhelmed
 Student was dismissed
 Student left the room and did not return
 Student has an accommodation that allows taking breaks
 Student was administered the assessment administration over multiple days

- Student refused to complete the assessment
- Environmental disruption resulted in student not completing the assessment

Other reason, please describe.

Was the student(s) provided an opportunity to participate in a practice session?

- All students were provided the opportunity
- Some students were provided the opportunity
- None of the students were provided the opportunity

Were any of the students or the specifically observed student observed choosing the same answer repeatedly?

- Yes
- No

If yes, was it related to any of the following?

- Test content
- Test preparation
- Student characteristic
- TA/Proctor/AA behavior
- Environment
- Unknown

Were any of the students or the specifically observed student observed hurrying through the assessment?

- Yes
- No

If yes, was it related to any of the following?

- Test content
- Test preparation
- Student characteristic
- TA/Proctor/AA behavior
- Environment
- Unknown

Were any of the students observed using the universal tools provided in the assessment?

- Yes
- No

If yes, how did the student appear to be using the tool(s)?

- Appropriately utilizing the tools
- Trying the tool out
- Playing around (tool appeared to be a distraction)
- Other, please describe.

List any observed accommodations provided to students.

Please provide any insight, including specific topics for additional assessment training offered by the Maine Department of Education.

Did the assessment platform function as expected?

Yes

No

If no, please describe and include what type of device was used (e.g., iPad, Chromebook, Windows).

Section 8: Achievement Standards and Reporting

Achievement standards are the descriptions defining student stands in the four achievement levels: *Well-Below State Expectations*, *Below State Expectations*, *At State Expectations*, and *Above State Expectations*. This section describes the procedures for defining achievement standards, setting achievement standards, and reporting.

8.1. State Adoption of Achievement Standards

The Maine Through Year Assessment (MTYA) program is Maine’s statewide system of summative assessments in reading and mathematics in grades 3–8 and the second year of high school that was first administered in Spring 2023. The Maine Department of Education (DOE) contracted with NWEA to design and develop the MTYAs, and NWEA contracted with edCount LLC and Creative Measurement Solutions LLC to design and implement the alignment study and standard setting.

The MTYA standard setting design is a systematic approach grounded in principled assessment design (PAD). Under this design, Achievement Level Descriptors (ALDs) are developed early in the test-development lifecycle to support domain definition (e.g., explication of the construct of interest), item development, and standard setting. Table 8.1 presents the four achievement levels established for the MTYA.

Table 8.1. MTYA Achievement Level Descriptors

Well-Below State Expectations	Below State Expectations	At State Expectations	Above State Expectations
On this assessment, students at this achievement level demonstrate limited understanding of the knowledge and skills necessary at this grade level, as specified in the Common Core State Standards. The students need substantial academic support to be prepared for the next grade level and to be on track for college and career readiness.	On this assessment, students at this achievement level demonstrate partial understanding of the knowledge and skills necessary at this grade level, as specified in the Common Core State Standards. The students need additional academic support to be prepared for the next grade level and to be on track for college and career readiness.	On this assessment, students at this achievement level demonstrate the knowledge and skills necessary at this grade level, as specified in the Common Core State Standards. The students are prepared for the next grade level and are on track for college and career readiness.	On this assessment, students at this achievement level demonstrate advanced understanding of the knowledge and skills necessary at this grade level, as specified in the Common Core State Standards. The students are well prepared for the next grade level and are well prepared for college and career readiness.

Three cut scores were adopted, defining the four levels of achievement:

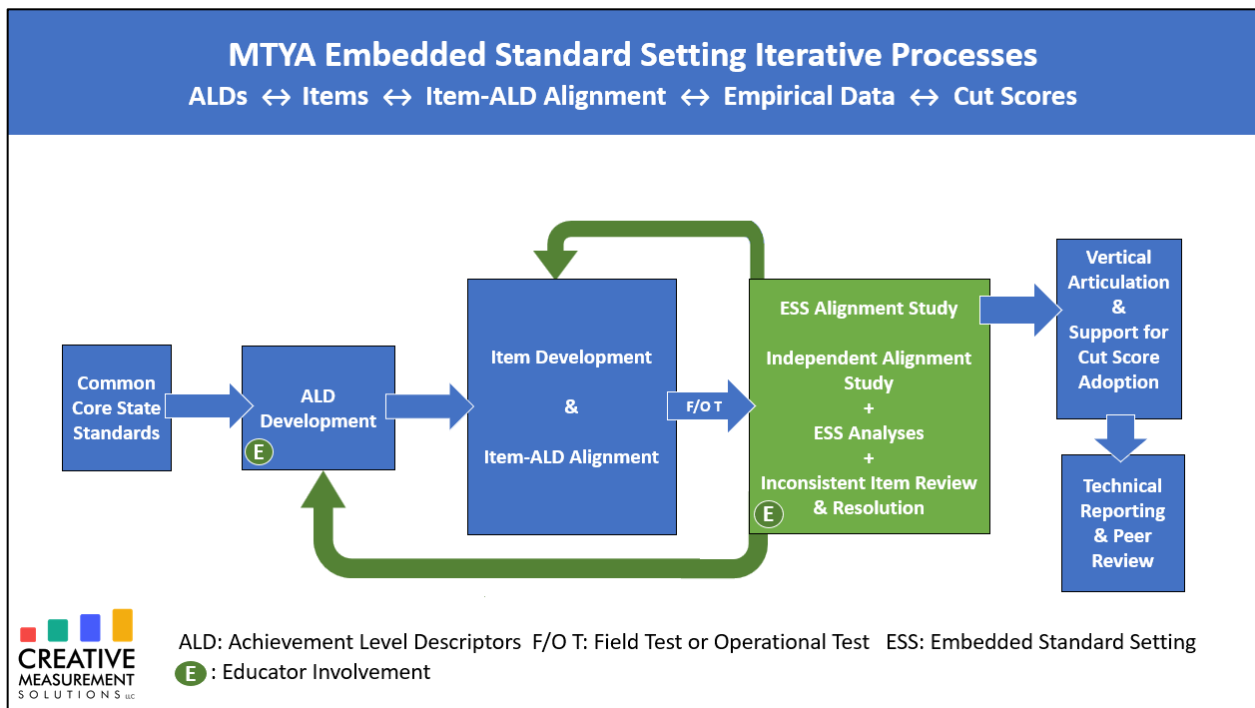
- The *Below State Expectations* cut score separates the *Well-Below State Expectations* and *Below State Expectations* levels.
- The *At State Expectations* cut score separates the *Below State Expectations* and *At State Expectations* levels.

- The *Above State Expectations* cut score separates the *At State Expectations* and *Above State Expectations* levels.

8.2. Achievement Standard Setting

Embedded Standard Setting (ESS) was employed to establish the Maine Through Year Assessment achievement level cut scores. The ESS methodology was selected because it is the natural extension of principled assessment design to standard setting (Lewis & Cook, 2020). It transforms standard setting from a standalone workshop to a set of processes actively integrated throughout the assessment-development lifecycle, as illustrated in Figure 8.1. The iterative nature of the ESS processes (represented by the green feedback arrows in the figure) supports the coherence of various assessment components and artifacts, including ALDs, item development, item-ALD alignment, empirical data, and cut scores (and, therefore, score interpretation). Thus, adherence to these iterative processes supports validity of the assessments and score interpretation.

Figure 8.1. Maine Through Year Assessment Embedded Standard Setting Iterative Processes



ESS processes directly contribute to the valid interpretation and use of test scores and improve test quality and the strength of validity arguments by maintaining a consistent focus on optimizing the evidentiary relationship between test items and the Common Core State Standards (CCSS), as reflected by the associated Achievement Level Descriptors (ALDs). ESS processes include:

- Achievement Level Descriptor development: an articulation of the intended interpretations of the Maine Through Year Assessment across the achievement levels.
- The ESS Alignment Study: a review of a representative sampling of MTYA items by Maine educators in which they provide independent alignments of these items to the

Common Core State Standards and Maine achievement levels and review and resolve items with alignments that are inconsistent with the data.

- ESS analyses and the estimation of cut scores: educators' alignments of items to the Maine achievement levels are employed to identify optimal cut scores.
- Post-ESS Alignment Study workshop: these activities lead to the adoption of cut scores, including cut score refinement to support an integrated, vertically articulated system of cross-grade cut scores meeting workshop panelists' and other stakeholders' expectations and in consideration of Maine DOE policy goals.
- Documentation of validity evidence supporting Maine's adopted cut scores: this includes those forms of evidence commonly cited in the measurement literature and those used to satisfy federal peer review requirements.

Findings from each of these activities provide evidence that the ESS processes work together to promote the coherence of the assessment. Specifically:

- Range ALDs were developed to align to the CCSS; final ALDs were reviewed and refined by Maine educators.
- Results from the ESS Alignment Study demonstrated the efficacy of panelists' consensus regarding the alignment of items to the ALDs; high correlations with empirical difficulty, weighted kappa values, and panelist agreement rates demonstrated a strong panelist understanding of their role and judgment tasks.
- ESS analyses produced cut scores that optimally reflect the panelists' judgments by minimizing inconsistencies between those judgments and empirical data.
- Results from the Review and Resolution workshop showed iterative improvement in the consensus regarding item-ALD alignments and associated efficacy measures, including correlations, kappa values, and agreement rates, as expected of a consensus-building activity.
- Post-workshop vertical articulation produced a well-articulated, cross-grade system of cut scores in mathematics and reading that reflect the panelists' and other stakeholders' expectations for impact data, using methods supported by MTYA Technical Advisory Committee members.
- Thorough documentation of validity evidence supporting the MTYA adopted cut scores demonstrated strong adherence to principles of test score validation, as articulated in the measurement literature and in the guidelines for federal peer review.

Together, these findings support the validity of the Maine Through Year Assessment program's adopted cut scores. Linkages from ALDs to test scores are consistent with the tenets of Principled Assessment Design, support intended score interpretations, and inform decision-making.

NWEA will present MDE with a recommended plan of action based on the results of the July 2023 alignment study. This plan will include a review of the ALD language as well as the knowledge, skills, and abilities for each during the progression from Level 1 to Level 4. NWEA will consult with MDE on the details of this plan as it works to finalize it.

For reading and mathematics, the adopted cut scores were presented to the Commissioner of Education and were approved on August 28, 2023. Table 8.2–Table 8.5 present the final approved cut scores that were used for scoring and the associated impact data.

Table 8.2. Final Approved Cut Scores—Reading

Grade	Cut Scores		
	<i>Below State Expectations</i>	<i>At State Expectations</i>	<i>Above State Expectations</i>
3	1483	1500	1525
4	1486	1500	1525
5	1487	1500	1525
6	1486	1500	1525
7	1483	1500	1525
8	1484	1500	1525
HS	1489	1500	1525

Table 8.3. Impact Data Associated with Cut Scores—Reading

Grade	Percent at Level			
	<i>Well-Below State Expectations</i>	<i>Below State Expectations</i>	<i>At State Expectations</i>	<i>Above State Expectations</i>
3	12.6%	27.1%	47.3%	13.0%
4	12.2%	23.9%	48.5%	15.4%
5	12.8%	18.6%	53.0%	15.6%
6	10.4%	22.5%	53.5%	13.6%
7	11.4%	24.9%	50.4%	13.3%
8	10.1%	24.2%	53.4%	12.3%
HS	13.3%	24.7%	49.7%	12.3%

Table 8.4. Final Approved Cut Scores—Mathematics

Grade	Cut Scores		
	<i>Well-Below State Expectations</i>	<i>Below State Expectations</i>	<i>At State Expectations</i>
3	1486	1500	1525
4	1488	1500	1525
5	1484	1500	1525
6	1481	1500	1525
7	1482	1500	1525
8	1484	1500	1525
HS	1489	1500	1525

Table 8.5. Impact Data Associated with Cut Scores—Mathematics

Grade	Percent at Level			
	<i>Well-Below State Expectations</i>	<i>Below State Expectations</i>	<i>At State Expectations</i>	<i>Above State Expectations</i>
3	17.3%	21.1%	43.9%	17.7%
4	18.6%	24.5%	44.0%	12.9%
5	18.5%	30.7%	40.0%	10.8%
6	18.8%	36.4%	35.9%	8.9%
7	20.1%	36.0%	35.4%	8.5%

Grade	Percent at Level			
	<i>Well-Below State Expectations</i>	<i>Below State Expectations</i>	<i>At State Expectations</i>	<i>Above State Expectations</i>
8	20.5%	39.1%	33.5%	6.9%
HS	25.0%	32.0%	35.5%	7.5%

8.3. Reporting

The Maine Through Year Assessments are administered in reading and mathematics. These assessments were developed specifically for Maine to provide teachers, students, and parents with information on student learning strengths and needs throughout the year, as well as student progress in mastering college and career-ready skills based on Maine’s accountability standards, the Common Core State Standards.

8.3.1. Achievement Level Descriptors

An achievement level is a range of scores that defines a specific level of student achievement, as articulated in the achievement level descriptors (ALDs). The ALDs are a plain-language description of what students must know as defined by each of the achievement levels established through cut scores. The ALDs firmly root the cut scores and achievement levels in the content that students are supposed to learn. In qualitative and quantitative terms, the ALDs and cut scores *together* define the difference between a student who is performing at, below, or above grade-level expectations.

- *Well-Below State Expectations*: On this assessment, students at this achievement level **demonstrate limited understanding of the knowledge and skills** necessary at this grade level, as specified in the Common Core State Standards. The students **need substantial academic support** to be prepared for the next grade level and to be on track for college and career readiness.
- *Below State Expectations*: On this assessment, students at this achievement level **demonstrate partial understanding of the knowledge and skills** necessary at this grade level, as specified in the Common Core State Standards. The students **need additional academic support** to be prepared for the next grade level and to be on track for college and career readiness.
- *At State Expectations*: On this assessment, students at this achievement level **demonstrate the knowledge and skills** necessary at this grade level, as specified in the Common Core State Standards. The students **are prepared** for the next grade level and are on track for college and career readiness.
- *Above State Expectations*: On this assessment, students at this achievement level **demonstrate advanced understanding of the knowledge and skills** necessary at this grade level, as specified in the Common Core State Standards. The students **are well prepared** for the next grade level and are well prepared for college and career readiness.

The cut scores for these achievement levels were established and validated in summer 2023 by Maine educators, the Maine DOE, and the Maine Technical Advisory Committee.

8.3.2. Setting the Cut Scores

To establish the cut scores, a process called “embedded standard setting” helps determine two points along the scale score range (known as cut scores) that define the

score range for each achievement level. Maine educators and stakeholders from around the state participated in the embedded standard-setting process for the Maine Through Year Assessment, facilitated by edCount and Creative Measurement. The cut score recommendations from this statewide committee were presented to the Maine Department of Education and were approved in late August 2023.

8.3.3. Reports

For the Maine Through Year Assessment, reports were developed and are available at the district, school, group, and individual student levels. Table 8.6 presents a description of each report. A more detailed report explanation can be found in Appendix F.

Table 8.6. Report Levels

Report Name	Aggregation Level	Summary
District Report	District	Shows the average scale scores for schools in the district, the distribution of school average scale scores across the achievement levels, and the distribution of student scale scores in each school.
School Report	School	Shows the average scale scores for students in the school, the distribution of student scale scores across the achievement levels, the average scale scores, score distributions for each group in the school, and the individual scale scores for each student in the school.
Teacher Report	Group	Shows the average scale scores for students in the group, the distribution of student scale scores across the achievement levels, and the individual scale scores for each student in the group.
Student Report	Individual student	Shows all the details for an individual student’s test.
Individual Student Report	Individual Student	Shows all tests in all available content areas for a student in this academic year. Designed for parents and families.
Demographic Report	Varies—based on user type	Shows the average scale scores, average reporting category scores, and distribution of scale scores for demographic groups such as gender, ethnicity/race, and targeted group.

Figure 8.2 and Figure 8.3 show a mockup of the Individual Student Report (ISR). The ISR is a two-page report that is designed to show a student’s achievement on the Maine Through Year reading and mathematics assessments to parents and families. Educators can print these reports in batches, making it easy to distribute after testing is complete. The Individual Student Reports are generated for the spring assessment and will not be available for the fall and winter assessments.

Figure 8.2. Individual Student Report—Page 1

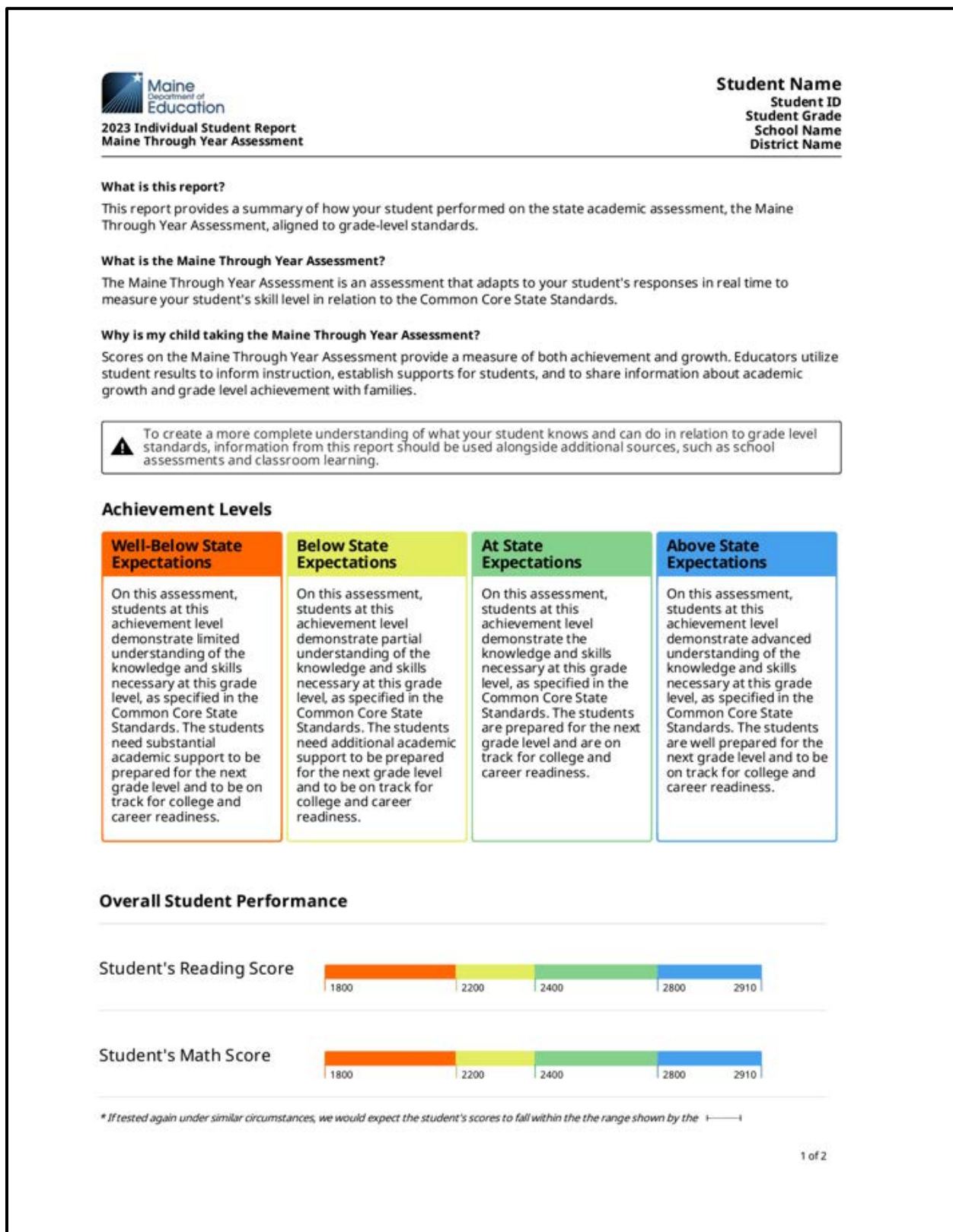
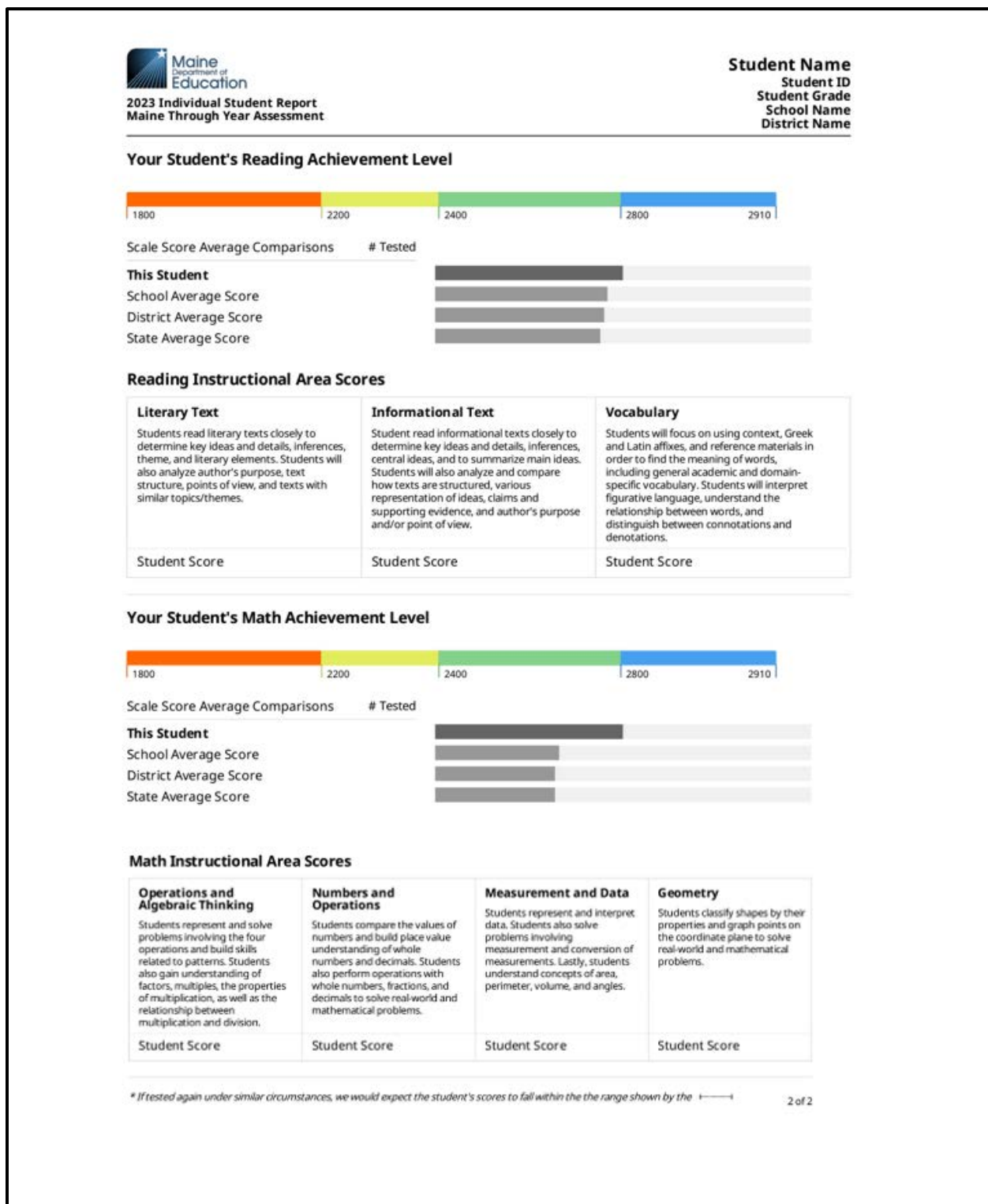


Figure 8.3. Individual Student Report—Page 2



For more report screenshots and report explanations, please see Appendix F.

Section 9: References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.
- Every Student Succeeds Act (ESSA), 20 U.S.C. § 6301 (2015).
<https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf>
- French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous Items. *Journal of Educational Measurement*, 33(3), 315–332.
<https://www.jstor.org/stable/1435375>
- Fu, J., & Monfils, L. (2016). *LDIF_ES: A SAS macro for logistic regression tests for differential item functioning of dichotomous and polytomous items*. (Research Memorandum ETS RM–16-17). Educational Testing Service (ETS).
<https://www.ets.org/Media/Research/pdf/RM-16-17.pdf>
- Gómez-Benito, J., Hidalgo, M. D., & Padilla, J.-L. (2009). Efficacy of effect size measures in logistic regression: An application for detecting DIF. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 5(1), 18–25.
<https://doi.org/10.1027/1614-2241.5.1.18>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer. <https://doi.org/10.1007/978-94-017-1988-9>
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer. <https://doi.org/10.1007/978-1-4939-0317-7>
- Lewis, D. & Cook, R. (2020). Embedded standard setting: Aligning standard setting methodology with contemporary assessment design principles. *Educational Measurement*, 39(1), 8–21. <https://doi.org/10.1111/emip.12318>
- Linacre, J. M. (2015). *Winsteps*® (Version 3.90.2) [Computer software]. Portland, Oregon: Winsteps.com. Available from <https://www.winsteps.com/>
- Linacre, J. M. (2002) What do infit and outfit, mean-square and standardization mean? *Archives of Rasch Measurement*, 16(2), 878. <https://www.rasch.org/rmt/rmt162f.htm>
- Linn, R. L. (2006). The standards for educational and psychological testing: Guidance in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development*. Routledge.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
<https://doi.org/10.1007/BF02296272>

- National Center for Research on Evaluation, Standards, & Student Testing (CRESST). (2015). *Simulation-based evaluation of the smarter balanced summative assessments*. [Tech. Rep.]. Retrieved from <https://portal.smarterbalanced.org/library/en/simulation-based-evaluation-of-the-smarter-balanced-summative-assessments.pdf>
- NWEA. (2020). *Constraint-based engine scientific approach and methodology* [Confidential Tech. Rep.].
- Phillips, S., & Camara, W. J. (2006). Educational measurement, 733–755.
- Puhan, G., Dorans, N. (2018). *Technical considerations in scale development*. Annual Meeting of the National Council on Measurement in Education, New York, NY, United States.
- Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press.
- Samajima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, 18(3), 229–244. <https://doi.org/10.1177/014662169401800304>
- Smarter Balanced Assessment Consortium (SBAC). (2016). *Smarter Balanced Assessment Consortium: 2014–15 technical report*. Retrieved from <https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf>.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370. <https://www.jstor.org/stable/1434855>
- Van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22(3), 259–270. <https://doi.org/10.1177/01466216980223006>
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97–116. <https://www.jstor.org/stable/1434010>
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zwick, R., Thayer, D. T., & Wingersky, M. (1994). DIF analysis for pretest items in computer-adaptive testing. (Research Report No. RR-94-33). Educational Testing Service (ETS). <https://doi.org/10.1002/j.2333-8504.1994.tb01606.x>