Maine
Department of
Education

# Maine Through Year Assessment
# Spring 2024 Technical Report

nwea

## Table of Contents

## List of Tables

## List of Figures

## Section 1: Overview of the Maine Through Year Assessment

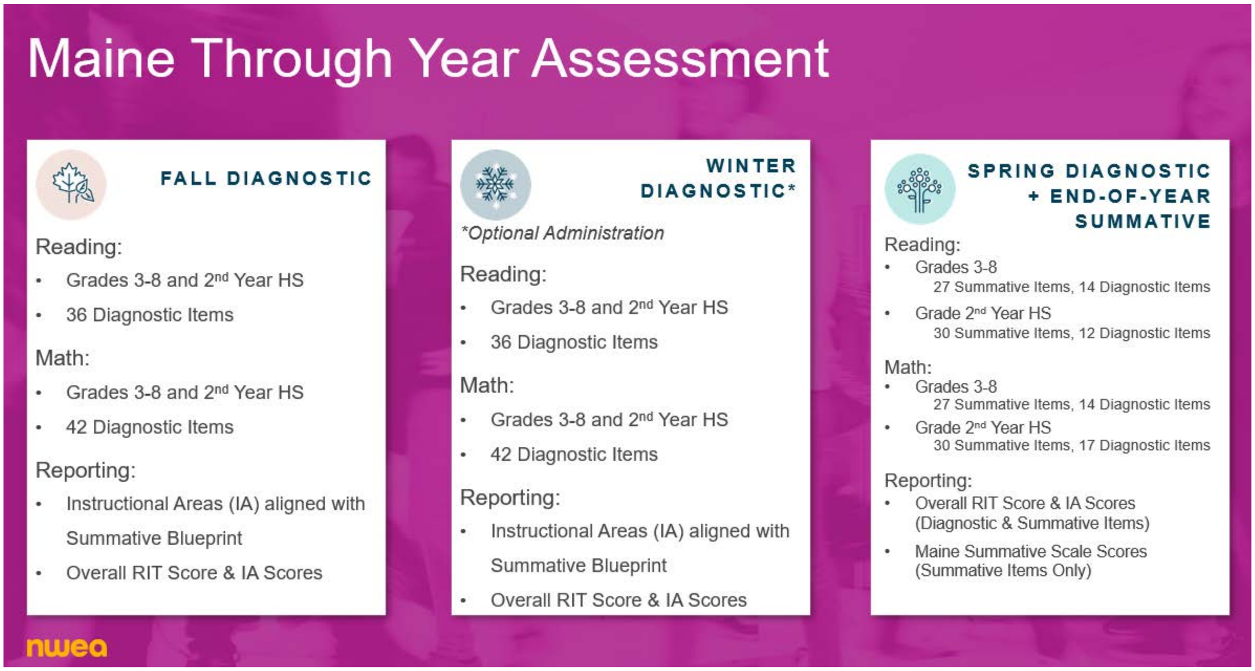
### 1.1. Structure of the Maine Through Year Assessment: A Balanced Assessment System

The Maine Through Year Assessment (MTYA) is a balanced assessment system that combines interim diagnostic assessments administered multiple times throughout the academic year with an end-of-year summative assessment. The interim assessments provide a measure of both student growth and achievement throughout the academic year and are separate from the end-of-year summative assessment, which reports student achievement according to grade-level state standards, specifically the Common Core State Standards.

The MTYA assesses all publicly funded Maine students in grades 3 through 8 and the second year of high school (HS/grade 10) in the content areas of reading and mathematics. The MTYA is 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.

The Through Year Assessment consists of three administrations: fall, winter, and spring. The fall and spring administrations are required for all students; the winter administration is optional. The fall and winter administrations are interim, diagnostic assessments 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 fulfills federal requirements for state assessments under the Every Student Succeeds Act (ESSA, 2015).

**Figure 1.1. Structure of the Maine Through Year Assessment**



Maine Through Year Assessment

FALL DIAGNOSTIC

Reading:
- Grades 3-8 and 2nd Year HS
- 36 Diagnostic Items

Math:
- Grades 3-8 and 2nd Year HS
- 42 Diagnostic Items

Reporting:
- Instructional Areas (IA) aligned with Summative Blueprint
- Overall RIT Score & IA Scores

WINTER DIAGNOSTIC*

*Optional Administration

Reading:
- Grades 3-8 and 2nd Year HS
- 36 Diagnostic Items

Math:
- Grades 3-8 and 2nd Year HS
- 42 Diagnostic Items

Reporting:
- Instructional Areas (IA) aligned with Summative Blueprint
- Overall RIT Score & IA Scores

SPRING DIAGNOSTIC + END-OF-YEAR SUMMATIVE

Reading:
- Grades 3-8
  27 Summative Items, 14 Diagnostic Items
- Grade 2nd Year HS
  30 Summative Items, 12 Diagnostic Items

Math:
- Grades 3-8
  27 Summative Items, 14 Diagnostic Items
- Grade 2nd Year HS
  30 Summative Items, 17 Diagnostic Items

Reporting:
- Overall RIT Score & IA Scores (Diagnostic & Summative Items)
- Maine Summative Scale Scores (Summative Items Only)

nwea

This technical report documents the processes and procedures implemented to support the end-of-year summative portion of the MTYA. For the purposes of this technical report, only the summative portion of the assessment will be discussed. 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 May 1, 2025.

The NWEA MAP® Growth$^{TM}$ Technical Report provides more information regarding elements of the diagnostic portions of the assessment, including item development and the computer-adaptive engine.

## 1.2. Intended Purposes and Uses of Test Results

The end-of-year summative portion of the MTYA 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, 2015) 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 summative portion of the MTYA is designed to measure Maine's accountability standards, the Common Core State Standards (CCSS), in mathematics and reading. Student results are reported according to academic achievement level descriptors utilizing cut scores established through embedded standard setting for each of the four achievement levels: *Well-Below State Expectations*, *Below State Expectations*, *At State Expectations*, *Above State Expectations*.

## 1.3. Required Assessment and Policies for Including All Students

All students in grades 3–8 and the second year of high school enrolled in Maine's public schools, Special Purpose Private Schools (SPPS), regional programs, charter schools, or private schools with at least 60% publicly funded students are required to participate in the MTYA. The participation requirement also applies to all students whose education is paid through Maine's public-school funds, even if those students are attending a private school. Publicly funded students are eligible for and required to participate in Maine's state assessment program at state expense. Students with disabilities and multilingual learners may participate in the MTYA with accommodations.

Exceptions to participation in the MTYA 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 MTYA.

## 1.4. Meaningful Consultation

### *1.4.1. Schedule of Major Events*

Table 1.1 presents the major events that occurred for the 2024 Maine Through Year Assessment.

**Table 1.1. Schedule of Major Events for the Spring 2024 Administration**

| Event | Date(s) |
|---|---|
| Content and Bias Review | August 8–9, 2023 |
| Technical Advisory Committee (TAC) Meeting | August 18, 2023 |
| Test Administration Training [a] | March 21 and 26, 2024 |
| Operational Test Window | April 22–May 31, 2024 |
| Data Review | October–November 2024 |

[a] Test Administration Training slides are included in Appendix A.

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

- Content and Bias Review: a meeting with Maine educators to review all items authored for the program by NWEA

- 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 H.

- Test Administration Training: training to prepare District Assessment Coordinators, School Assessment Coordinators, and proctors. Topics covered include Maine Through Year Assessment Overview, Technology Readiness, Assessment Management in Acacia, Accessibility & Not-Tested Codes, Preparing & Monitoring the Assessment, Proctor & Student Experience, Operational Reports, Data & Reporting, Preparation, Resources, & Tips, and Communication & 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 Maine DOE the results of this review, and decisions are made regarding the next steps for the flagged items.

Below is a list of topics from the TAC meeting leading up to the Spring 2024 Through Year Assessment administration:

- August 18, 2023
  - Calibration Results
  - Scaling Method
  - Technical Report Template Review and Discussion
  - Comparability Evidence
  - Score Comparisons
  - Standard Setting Technical Report
  - Embedded Standard Setting and Alignment Study Discussion and Cut Score Review and Discussion

# Section 2: Test Design and Content Development

This section describes the test design and content development processes for the Spring 2024 Maine Through Year Assessment.

## 2.1. Test Design & Development

The Maine Through Year Assessment (MTYA) is designed to measure Maine's accountability standards, the Common Core State Standards (CCSS), in mathematics and reading. In Spring 2024, Maine administered computer adaptive assessments in reading and mathematics for grades 3–8 and a limited adaptive assessment in reading and mathematics for the second year of high school (HS/grade 10). The reading and mathematics HS assessments were fixed forms in Spring 2023; the Spring 2024 assessments were adaptive tests with shallow item pools and included a focus on field testing to expand the operational pool and support increased adaptability in Spring 2025. The summative items in the grades 3–8 and HS assessments were licensed from NWEA; the grades 3–8 items came from existing NWEA item banks, and the HS items were part of a bank created by NWEA in collaboration with the Maine DOE and Maine educators. All items on the assessment were aligned to the CCSS and underwent a rigorous item-development process.

Table 2.1 summarizes the versions of the assessments available in Spring 2024. The paper forms were available in three formats (standard print, large print, or braille) and did not include field test items. While only the summative portions of the assessments are covered in this technical report, the table includes diagnostic item counts and overall test length for the spring assessment.

**Table 2.1. Summary of Assessments Available by Content Area & Grade for Spring 2024**

| Grade | Summative Items | Approximate Summative Points | Field Test Items | Diagnostic Items | Total Items | Paper Form? |
|---|---|---|---|---|---|---|
| **Reading** | | | | | | |
| 3 | 27 | 30–31 | 5 | 14 | 46 | Yes |
| 4 | 27 | 30–31 | 5 | 14 | 46 | Yes |
| 5 | 27 | 30–31 | 5 | 14 | 46 | Yes |
| 6 | 27 | 30–31 | 5 | 14 | 46 | Yes |
| 7 | 27 | 30–31 | 5 | 14 | 46 | Yes |
| 8 | 27 | 30–31 | 5 | 14 | 46 | Yes |
| HS | 30 | 33–34 | 7 | 12 | 49 | Yes |
| **Mathematics** | | | | | | |
| 3 | 27 | 30–31 | 5 | 18 | 50 | Yes |
| 4 | 27 | 30–31 | 5 | 18 | 50 | Yes |
| 5 | 27 | 30–31 | 5 | 18 | 50 | Yes |
| 6 | 27 | 30–31 | 5 | 18 | 50 | Yes |
| 7 | 27 | 30–31 | 5 | 18 | 50 | Yes |
| 8 | 27 | 30–31 | 5 | 18 | 50 | Yes |
| HS | 30 | 33–34 | 5 | 17 | 52 | Yes |

## 2.2. Test Blueprints

All items on the end-of-year summative portion of the MTYA are aligned to a Common Core State Standard and to an achievement level descriptor specific to the standard. To ensure coverage of grade-level academic standards, the Maine Through Year summative test blueprints for reading and mathematics outline the overall structure of the assessments. For Maine, the summative blueprints are structured around instructional areas for reporting based on content categories within the CCSS.

The assessment blueprints for each grade level in reading and mathematics in Appendix G list:

- the grade-level content standards included within each instructional area,
- the item-count targets for each instructional area,
- the approximate points targets and approximate percentage-of-overall-points targets on the assessment for each instructional area,
- the percentage of on-grade items, and
- the achievement level percentage targets.

Appendix B provides more detailed information about standard coverage at each grade level within the blueprints based on empirical data from the Spring 2024 administration.

### *2.2.1. Cognitive Complexity Blueprint Considerations*

Cognitive complexity should be considered as part of the guidelines for constructing the test blueprints. The MTYA blueprints include targets specifying boundaries for the percentages of items that should be selected from the different achievement levels, with at least 60% or more of the items coming from *At State Expectations* and *Above State Expectations* levels. Because the MTYAs in reading and mathematics are adaptive, the exact distribution of Achievement Level Descriptors (ALDs) and ALD levels for any given test event will vary based on individual student achievement and other blueprint constraints.

For the MTYA, each item in the reading and mathematics pools was either written for or aligned to a specific standard and ALD representing the skills and cognitive processing complexity of the item. To ensure the assessments include a deep pool of items that span a full range of cognitive levels and skills, both the standards and the ALD distributions are important factors when considering item pool needs and item-development plans.

### *2.2.2. Mathematics Summative Blueprint Considerations*

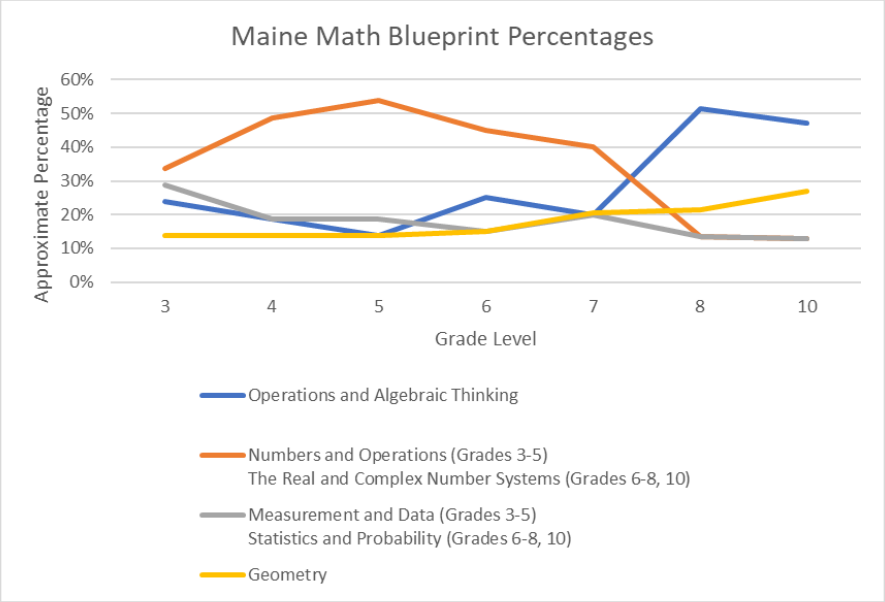
For mathematics, the instructional areas are closely connected to the CCSS mathematics domains, as shown in Table 2.2.

**Table 2.2. Instructional Areas for Maine Mathematics Summative Blueprints**

| **Instructional Areas for Grades 3 to 5** |
|---|
| Operations and Algebraic Thinking |
| Numbers and Operations |
| Measurement and Data |
| Geometry |
| **Instructional Areas for Grades 6 to 8 and HS** |
| Operations and Algebraic Thinking |
| The Real and Complex Number Systems |
| Geometry |
| Statistics and Probability |

The mathematics blueprints reflect the instructional emphasis of the content at each grade. For example, Geometry receives more instructional time as the grades progress, which is reflected in how the percent increases from 10% in grade 3 to 30% in grade 10 (HS). The blueprints were also influenced by Student Achievement Partners' Focus Areas in Mathematics (Achieve the Core), which calls out the major work of each grade and guides educators on how to focus instructional time. Figure 2.1 shows how the percentage for each reporting category shifts across the grades.

**Figure 2.1. Maine Blueprint Percentages—Mathematics, Grades 3–8 & 10 (HS)**



The graph shows that students' skills in the Numbers and Operations instructional area are emphasized as they work with whole numbers less than 1,000 and fractions with a limited set of denominators in grade 3 before moving on to decimals and a larger set of fractions in grade 5. After students grasp these skills, the significance of the instructional area (which shifts to The Real and Complex Number Systems starting in grade 6) gradually lessens as students work with the set of rational numbers in grade 6 before moving on to the set of irrational numbers in high school.

Conversely, the emphasis on students' skills in the Operations and Algebraic Thinking instructional area steadily increases as students solve simple two-step problems in context in grade 3 and move to working with linear and quadratic functions in high school.

For the content category Measurement and Data in grades 3–5 and Statistics and Probability in grades 6–8 and HS, the emphasis remains relatively constant and ranges from 10% to 30%. Students' skills gradually progress from working with picture graphs in grade 3 to scatter plots in high school.

For the Geometry instructional area, the emphasis gradually increases from around 14% in grade 3 as students work with area and perimeter to around 28% in high school as students work with more complex figures and geometric proofs.

In grade 7, three content categories are each assessed at approximately 20% in the blueprint because it is the point at which the Operations and Algebraic Thinking instructional area and the Geometry instructional area continue to increase while the third instructional area (Measurement and Data in grades 3–5 and Statistics and Probability in grades 6–8 and HS) remains relatively constant near 20%.

Table 2.3 and Table 2.4 show the approximate percentages for the instructional areas for each grade. Additional information about how these percentages are represented in the assessments can be found in Appendix G.

**Table 2.3. Approximate Summative Blueprint Percentages: Mathematics, Grades 3–5**

| Instructional Area | 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% | 18–20% | 18–20% |
| Geometry | 13–15% | 13–15% | 13–15% |

**Table 2.4. Approximate Summative Blueprint Percentages: Mathematics, Grades 6–8 & HS**

| Instructional Area | Grade 6 | Grade 7 | Grade 8 | HS |
|---|---|---|---|---|
| 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*

In reading, the instructional areas shown in Table 2.5 are closely connected to the CCSS Reading Strands.

**Table 2.5. Instructional Areas for Maine Reading Summative Blueprints**

| Instructional Areas for Grades 3–8 & HS |
|---|
| Literary Text |
| Informational Text |
| Vocabulary |

When creating the reading blueprints for Maine, focus was given to the weight and breadth of the reading standards designed to assess literary and informational texts and vocabulary skills. The blueprints were influenced by the Priority Instructional Content guidance from Student Achievement Partners (Achieve the Core) and reflect the belief that not all content standards are emphasized equally in the classroom. The reading assessments are designed to keep the text at the center and use text-based questions. These assessment items highlight close reading skills, text analysis, textual evidence, and academic vocabulary.

Table 2.6 shows the approximate percentages for the instructional areas for each grade. Additional information about how these percentages are represented in the assessments can be found in Appendix G.

**Table 2.6. Approximate Blueprint Percentages by Instructional Area: Reading, Grades 3–8 & HS**

| Instructional Area | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | HS |
|---|---|---|---|---|---|---|---|
| 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% |

It is important that reading assessments provide a balance of content between literary and informational texts and represent a range of text complexity. According to the Common Core State Standards, students are expected to demonstrate an 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 2.7 shows the percentages of literary and informational text by grade.

**Table 2.7. Approximate Blueprint Percentages by Text Type: Reading, Grades 3–8 & HS**

| Text Type | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 | HS |
|---|---|---|---|---|---|---|---|
| Literary | 55–60% | 55–60% | 50% | 45–50% | 40–45% | 40–45% | 40–45% |
| Informational | 40–45% | 40–45% | 50% | 45–50% | 55–60% | 55–60% | 55–60% |

Text complexity is a measure of how challenging a text is to read and understand. Many factors may make a text complex, so a text complexity measurement is the process of evaluating a text for quantitative data, qualitative data, and the considerations for the reader and task. The texts in the reading item bank should include those that cover a range of text complexity within a grade level, including 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 including Lexile, word count, and Flesch-Kincaid. Quantitative measures are only a guide, exceptions can be made if the qualitative measures and/or grade-level alignments are appropriate. Table 2.8 shows acceptable Lexile ranges for each grade.

**Table 2.8. Approximate Reading Lexile Ranges, Grades 3–8 & HS**

| Grade(s) | Lexile Range |
|---|---|
| 3 | 450L–790L |
| 4–5 | 745L–980L |
| 6–8 | 925L–1155L |
| HS | 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 2.9 provides acceptable word count ranges for each grade. For paired passages, each individual passage should fall within the word count range.

**Table 2.9. Approximate Reading Word Count Ranges, Grades 3–8 & HS**

| Grade | Word Count Range |
|---|---|
| 3 | 200–700 |
| 4 | 200–900 |
| 5 | 300–1000 |
| 6 | 400–1100 |
| 7 | 400–1100 |
| 8 | 400–1200 |
| HS | 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 using a Passage Quality Checklist (included in Appendix J) that documents the complexity and suitability of each passage for assessment. For more information about text complexity, see https://achievethecore.org/page/2725/text-complexity.

## 2.3. Item Types

The Maine Through Year Assessment consists of several item types, as outlined in Table 2.10.

**Table 2.10. Online Item Types**

| Item Type | Description |
|---|---|
| Multiple-Choice (Choice) | Students select one response from multiple options. |
| Multi-Select (Choice Multiple) | Students select two or more responses from multiple options. Some multi-select 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 2.11 and Table 2.12 outline the percentages of item types by content area and grade level in the available Spring 2024 summative pools.

**Table 2.11. Item Type Percentages by Grade—Reading Summative Pools, Spring 2024**

| Grade | Item Type | | | | |
|---|---|---|---|---|---|
| | Multiple-Choice | Multi-Select | Composite | Gap Match | Hot Text |
| 3 | 79% | 8% | 6% | 6% | 1% |
| 4 | 82% | 8% | 5% | 5% | 0% |
| 5 | 84% | 8% | 2% | 4% | 2% |
| 6 | 84% | 6% | 4% | 4% | 2% |
| 7 | 76% | 11% | 6% | 6% | 1% |
| 8 | 84% | 6% | 4% | 5% | 2% |
| HS | 63% | 14% | 18% | 4% | 0% |

**Table 2.12. Item Type Percentages by Grade—Mathematics Summative Pools, Spring 2024**

| Grade | Item Type | | | | | | |
|---|---|---|---|---|---|---|---|
| | Multiple-Choice | Multi-Select | Composite | Gap Match | Graphic Gap Match | Hot Text | Text Entry |
| 3 | 53% | 10% | 6% | 10% | 6% | 3% | 12% |
| 4 | 52% | 11% | 4% | 8% | 9% | 7% | 10% |
| 5 | 53% | 8% | 7% | 7% | 8% | 4% | 12% |
| 6 | 56% | 7% | 6% | 9% | 3% | 6% | 13% |
| 7 | 62% | 8% | 3% | 7% | 2% | 8% | 11% |
| 8 | 52% | 8% | 5% | 7% | 4% | 11% | 15% |
| HS | 33% | 22% | 16% | 14% | 2% | 11% | 2% |

## 2.4. 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). Maine's policy ALDs were adopted in Spring 2023 to broadly define the characteristics of student performance on the state assessment at each of the four achievement levels: *Well Below*, *Below*, *At*, and *Above State Expectations*. Table 2.13 provides detailed explanations of each policy ADL.

**Table 2.13. Maine's Policy 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 | 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 | 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 | On this assessment, students at this achievement level **demonstrate advanced understanding of the knowledge and skills** necessary at this grade level, as |

| ***Well-Below State Expectations*** | ***Below State Expectations*** | ***At State Expectations*** | ***Above State Expectations*** |
|---|---|---|---|
| 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. | 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. | Standards. The students ***are prepared*** for the next grade level and are on track for college and career readiness. | 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. |

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 based on the broader policy ALDs. 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. They do this by focusing on differentiating factors within each standard that represent the progression of student knowledge and understanding of the specified skill. The ALDs also strive to preserve differentiation between the skills as they progress across grades. 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, CCSS mathematics 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. By helping to describe a student's current level of understanding, range ALDs support stakeholders in pinpointing areas of strength and areas of growth. Range ALDs are also used to guide NWEA content specialists in writing items for assessments and, by doing so, help develop a deeper item bank that can better serve the needs of each individual student.

NWEA content specialists wrote the initial draft of the Maine range ALDs and then held a workshop with Maine educators in September 2022 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 HS in either reading or mathematics. The review's purpose was to allow Maine educators to study the ALDs and share their feedback with NWEA content specialists.

The number of committee members for each content area was limited, and for this reason, educators with expertise in all grade levels were recruited to participate. The four selected 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, and some participants had experience working with special education students, English language learners, and gifted and talented students.

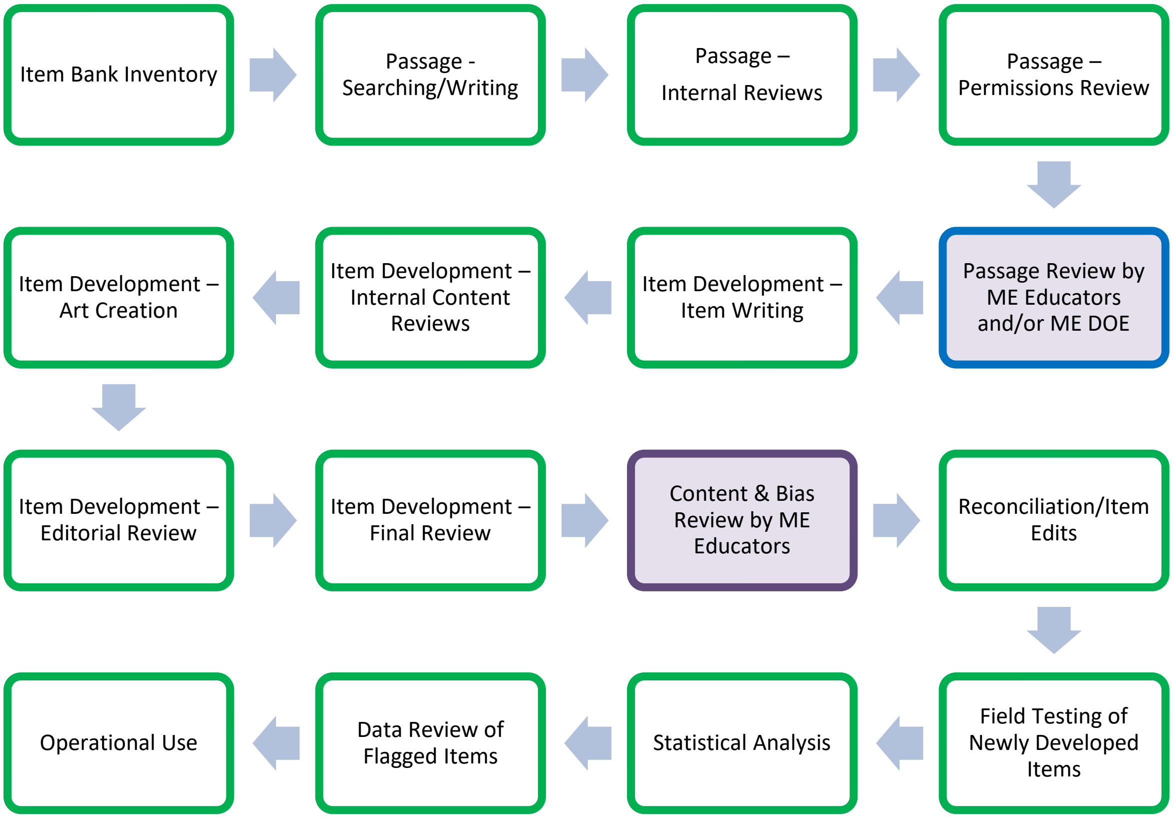
Both the reading and mathematics ALDs had progressions updated based on feedback from the Maine educators. These updates included reassigning ALD statements to another level within the progression, removing ALD statements, revising ALD statements, and crafting new ALD statements.

The complete set of range ALD statements utilized for the Maine Through Year Assessment is publicly available within the Achievement Level Explorer tool.

### 2.5. Content Development

New content development for Maine was focused on reading and mathematics HS items; however, the general content development process also applies to the grades 3–8 items that were selected for use in Maine from existing NWEA item banks. Items are developed in accordance with Universal Design for Learning principles and are each aligned to a standard and an ALD. Items are rigorously reviewed and edited during internal reviews, including reviews for bias/sensitivity at each stage of development. For the newly developed HS items, feedback from an external content and bias/sensitivity review involving Maine educators was incorporated into the items prior to field testing. Items that pass the review stages are field tested and subsequently reviewed based on their item statistics, which may include re-examining the item content. Items that pass the data review become operational in the item pool. Figure 2.2 provides an example of an item development workflow for reading; the mathematics workflow is similar but excludes the passage steps. At the conclusion of a test window, the process begins again, using the most recent pool and test simulation data to determine the areas of focus for future item development, with a particular focus on standard and ALD coverage.

**Figure 2.2. Reading Development Workflow**


Item Bank Inventory
Passage - Searching/Writing
Passage – Internal Reviews
Passage – Permissions Review
Passage Review by ME Educators and/or ME DOE
Item Development – Item Writing
Item Development – Internal Content Reviews
Item Development – Art Creation
Item Development – Editorial Review
Item Development – Final Review
Content & Bias Review by ME Educators
Reconciliation/Item Edits
Field Testing of Newly Developed Items
Statistical Analysis
Data Review of Flagged Items
Operational Use

### *2.5.1. Item Development and Guidelines*

Item development begins with a review and inventory of the existing item bank while also determining areas of focus based on information derived from simulations. From this, an item development plan can be created that may include item-writing specifications and targets such as:

- Specific standards to be targeted to strengthen and add more depth to the pool
- Specific cognitive complexity targets (in the form of ALDs) to strengthen and add more depth to the pool
- General item-writing guidelines in terms of overall content, item stems, item responses, style, and scoring rules
- Guidelines for using technology-enhanced items (TEIs)

The reading and mathematics HS items are written internally by NWEA content specialists or external professional item writers. Grades 3–8 include items written by NWEA content specialists, external professional item writers, and educators trained at item writing workshops. Regardless of the source, all items undergo a rigorous review and editing process, including bias/sensitivity reviews at multiple stages of development, and are developed in accordance with Universal Design for Learning principles. Following best practices, including style, ensures that items are accurately measuring student knowledge at each level by focusing the items on construct-relevant information and presentation and ensuring that items are accessible and fair to all students. The subsequent field-testing process, statistical analysis of item performance, and data review help further ensure the quality of items that become operational. All summative operational items in the MTYA, regardless of their original source, go through Maine-specific field testing to ensure they are effective items for Maine students.

For reading items, this process also includes writing or identifying passages and the identification or development of passage resources. Passage resources may include sources from the public domain, copyright works that are permissioned for use, and commissioned works. For HS item development, all passages are commissioned or taken from the public domain. Passages, like items, undergo a rigorous review process, including bias/sensitivity reviews.

Passages are developed or selected 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 and ALDs

Items developed for the Maine assessments are tightly aligned to either a part of or an entire standard as well as to a corresponding ALD for the standard, providing additional information around the cognitive demand and rigor of each item. Additional data points may also be tracked, such as Webb's depth of knowledge level, to support overall alignment decisions. Examples of the item and passage review checklists used during the content development process can be found in Appendix J.

### *2.5.2. 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 minimize the amount of background knowledge needed to correctly respond to an item and ensures that items do not contain sources of construct-irrelevant variance so that assessments can more appropriately capture what each student knows. Including UDL principles in item development also helps ensure that there will be available items for the creation of accommodated forms such as large print and braille.

### *2.5.3. Sensitivity and Fairness*

NWEA takes seriously the task of creating items that are free from bias and sensitivity issues and are fair to all students. Items are revised to eliminate bias, sensitivity, and fairness issues—or rejected if an issue cannot be remedied through the revision process. Items are reviewed for sensitivity and fairness multiple times throughout the development process, with some items also being reviewed in a collaborative effort with Maine educators (see Section 2.6).

- **Bias:** This is defined as item content, unrelated to the concept or skill being assessed, that may unfairly influence a student's performance or an item construct that does not have equivalent meaning for all students.
- **Sensitivity:** This can result if the experience of taking a test differs from the classroom experience in that students do not have the opportunity to discuss the material with a teacher or their peers. Sensitive content risks drawing students out of the testing experience by provoking negative emotional responses.
- **Fairness:** This is defined as the equitable treatment of all students during the assessment process. To make a test fair, test developers must work to eliminate any barriers that prevent students from understanding and interacting with item content in a manner that accurately demonstrates what they know or are able to do.

A successful item is free of bias and sensitivity issues and is accessible to all students. An item should NOT:

- Distract, upset, or confuse in any way
- Contain inappropriate or offensive topics
- Require construct-irrelevant knowledge or specialized knowledge
- Favor students from certain language communities
- Favor students from certain cultural backgrounds
- Favor students based on gender
- Favor students based on social economic issues
- Employ idiomatic or regional phrases and expressions
- Stereotype certain groups of people or behaviors
- Favor students from certain geographic regions
- Favor students who have no visual impairments
- Use height, weight, test scores, or homework scores as content or data in an item

There is no hard and fast "list" of material that is potentially distracting or upsetting, but some topics are seldom appropriate for K–12 assessments, such as sexuality, illegal substances, illegal activities, excessive violence, discriminatory descriptions, death, grieving, catastrophes, animal neglect or abuse, and loss of a family member.

## 2.6. Content and Bias Review Meeting

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 to gain actionable feedback on items. Only the high school items went through a CBR meeting in August 2023, as this was the only grade for which items were developed specifically for Maine. Grades 3–8 items, while not part of the Maine CBR process, would have undergone similar reviews either internally or externally during their initial development.

Educators are asked to review the items in advance of the virtual CBR and decide if they feel the items should be accepted, accepted with revisions, or rejected. 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. Following the general session, participants report to either the reading or mathematics breakout room, where reminders about the criteria by which items should be reviewed are provided.

Each breakout room includes an NWEA facilitator who leads a discussion regarding any items that have been flagged as accept with revisions or rejected by any of the reviewers with the goal of coming to a final decision for the item. If needed, requests are reconciled with Maine DOE in the days following the meeting and then revisions can be applied.

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

- Item aligns to the standards.
- Item is clearly worded.
- Item type is appropriate for the content/standard.
- Item has one and only one best correct answer.
- Item distractors are plausible.
- Item art is clear and necessary.
- Item is mathematically correct.
- Item is factually 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 2.14 outlines the total numbers of items taken to the Content and Bias Review meeting, as well as the numbers of items accepted, accepted with revisions, and rejected.

**Table 2.14. August 2023 Content and Bias Review Results**

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

### 2.7. Field Testing and Data Review

Data review is the process of reviewing field-tested items for quality and appropriateness based on the results of statistical analysis of student responses. In performing this data review, NWEA adheres to the *Standards for Educational and Psychological Testing* (AERA et al., 2014) 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 July/August 2024. 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 2025. Table 2.15 presents the criteria for flagging items that were field tested in Spring 2024, and Table 2.16 provides the results of the data review process.

**Table 2.15. Data Review Flagging Criteria—Multiple-Choice and Non-Multiple-Choice Items**

| Type | Label | Statistic | Flag |
|---|---|---|---|
| MC items | Pvalue_LOW/ Pvalue_HIGH | P value | < 0.2 or > 0.9 |
| | Pvalue_Dis | Option percentages | Distractor %<br>> p 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 | P 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 ><br>Step 2 |
| DIF | itemFlag_Gender_DIF/<br>itemFlag_Black_DIF/<br>itemFlag_Hispanic_DIF | DIF of gender or ethnicity | C+ or C- |

**Table 2.16. Data Review Results**

| Grade | Promote to OP | Revise and Re-Field Test | Reject | Total |
|---|---|---|---|---|
| **Mathematics** | | | | |
| 3 | 5 | 1 | 0 | 6 |
| 4 | 11 | 0 | 0 | 11 |
| 5 | 5 | 0 | 0 | 5 |
| 6 | 6 | 0 | 0 | 6 |
| 7 | 7 | 0 | 0 | 7 |
| 8 | 5 | 1 | 0 | 6 |
| HS | 143 | 2 | 0 | 145 |
| **Total** | **182** | **4** | **0** | **186** |
| **Reading** | | | | |
| 3 | 5 | 0 | 0 | 5 |
| 4 | 1 | 3 | 1 | 5 |
| 5 | 3 | 3 | 1 | 7 |
| 6 | 4 | 1 | 0 | 5 |
| 7 | 5 | 0 | 0 | 5 |
| 8 | 7 | 4 | 1 | 12 |
| HS | 110 | 4 | 3 | 117 |
| **Total** | **135** | **15** | **6** | **156** |

# 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 2024 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 3.4 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 | Hispanic/ Latino | Am. Indian or Alaska Native | Asian | Black or African American | Native HI or Pacific Islander | White | Two or More Races |
| 3 | N | 11,670 | 5,664 | 6,006 | 383 | 81 | 137 | 589 | 11 | 10,001 | 463 |
| | % | 100 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 4 | N | 12,237 | 5,985 | 6,252 | 443 | 81 | 150 | 569 | 16 | 10,490 | 487 |
| | % | 100 | 49 | 51 | 4 | 1 | 1 | 5 | 0 | 86 | 4 |
| 5 | N | 12,214 | 5,882 | 6,332 | 413 | 108 | 169 | 576 | 14 | 10,467 | 464 |
| | % | 100 | 48 | 52 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 6 | N | 11,953 | 5,833 | 6,120 | 404 | 97 | 154 | 534 | 13 | 10,292 | 457 |
| | % | 100 | 49 | 51 | 3 | 1 | 1 | 4 | 0 | 86 | 4 |
| 7 | N | 12,207 | 6,013 | 6,194 | 397 | 86 | 163 | 604 | 6 | 10,503 | 451 |
| | % | 100 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 8 | N | 12,292 | 5,877 | 6,415 | 410 | 93 | 158 | 595 | 12 | 10,597 | 429 |
| | % | 100 | 48 | 52 | 3 | 1 | 1 | 5 | 0 | 86 | 3 |
| HS | N | 12,511 | 5,988 | 6,523 | 410 | 94 | 235 | 584 | 11 | 10,764 | 428 |
| | % | 100 | 48 | 52 | 3 | 1 | 2 | 5 | 0 | 86 | 3 |

**Table 3.2. Demographic Information—Mathematics**

| Grade | Type | Total | Gender | | Ethnicity | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Female | Male | Hispanic/ Latino | Am. Indian or Alaska Native | Asian | Black or African American | Native HI or Pacific Islander | White | Two or More Races |
| 3 | N | 11,723 | 5,695 | 6,028 | 395 | 79 | 141 | 619 | 11 | 10,007 | 466 |
| | % | 100 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 85 | 4 |
| 4 | N | 12,290 | 6,010 | 6,280 | 451 | 81 | 151 | 592 | 16 | 10,508 | 490 |
| | % | 100 | 49 | 51 | 4 | 1 | 1 | 5 | 0 | 86 | 4 |
| 5 | N | 12,259 | 5,904 | 6,355 | 420 | 106 | 170 | 604 | 14 | 10,477 | 465 |
| | % | 100 | 48 | 52 | 3 | 1 | 1 | 5 | 0 | 85 | 4 |
| 6 | N | 11,993 | 5,849 | 6,144 | 415 | 98 | 155 | 554 | 13 | 10,300 | 456 |
| | % | 100 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 7 | N | 12,237 | 6,038 | 6,199 | 400 | 86 | 165 | 621 | 6 | 10,511 | 452 |
| | % | 100 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |

| Grade | Type | Total | Gender | | Ethnicity | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Female | Male | Hispanic/ Latino | Am. Indian or Alaska Native | Asian | Black or African American | Native HI or Pacific Islander | White | Two or More Races |
| 8 | N | 12,335 | 5,897 | 6,438 | 422 | 94 | 159 | 611 | 12 | 10,606 | 433 |
| | % | 100 | 48 | , | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| HS | N | 12,551 | 6,010 | 6541 | 423 | 91 | 239 | 599 | 11 | 10,771 | 432 |
| | % | 100 | 48 | 52 | 3 | 1 | 2 | 5 | 0 | 86 | 3 |

**Table 3.3. Ability Distribution—Summative Scale Scores**

| Grade | Summative Theta | | | |
|---|---|---|---|---|
| | Reading | | Mathematics | |
| | Mean | SD | Mean | SD |
| 3 | 1504 | 16 | 1505 | 20 |
| 4 | 1505 | 16 | 1502 | 17 |
| 5 | 1507 | 16 | 1501 | 17 |
| 6 | 1507 | 15 | 1496 | 19 |
| 7 | 1507 | 16 | 1496 | 18 |
| 8 | 1505 | 17 | 1496 | 16 |
| HS | 1504 | 15 | 1500 | 18 |

**Table 3.4. Ability Distribution—Summative Theta**

| Grade | Summative Theta | | | |
|---|---|---|---|---|
| | Reading | | Mathematics | |
| | Mean | SD | Mean | SD |
| 3 | -0.29 | 1.34 | -0.54 | 1.59 |
| 4 | 0.11 | 1.27 | -0.47 | 1.72 |
| 5 | 0.14 | 1.23 | -0.09 | 1.65 |
| 6 | 0.16 | 1.13 | -0.46 | 1.64 |
| 7 | 0.33 | 1.22 | -0.87 | 1.69 |
| 8 | 0.38 | 1.34 | -0.76 | 1.50 |
| HS | 0.14 | 0.97 | -1.30 | 1.02 |

*3.2.2. Item Pool Characteristics*

To ensure the adequacy of the item pool for administering a computer adaptive test (CAT), Table 3.5 details the numbers of items of various types and levels in the item pool for Maine by instructional area in the summative item pools for reading and mathematics.

**Table 3.5. Numbers of Items by Content and Instructional Areas**

| Content Area | Instructional Area | Grade | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 3 | 4 | 5 | 6 | 7 | 8 | HS |
| Reading | Informational Text | 138 | 181 | 181 | 190 | 233 | 215 | 32 |
| | Literary Text | 162 | 186 | 177 | 149 | 132 | 116 | 8 |
| | Vocabulary | 83 | 126 | 139 | 111 | 134 | 133 | 9 |
| | **Total** | **383** | **493** | **497** | **450** | **499** | **464** | **49** |
| Mathematics | Geometry | 40 | 50 | 78 | 55 | 84 | 138 | 21 |
| | Measurement and Data | 172 | 105 | 84 | – | – | – | – |
| | Numbers and Operations | 148 | 255 | 269 | – | – | – | – |
| | Operations and Algebraic Thinking | 138 | 81 | 67 | 127 | 96 | 194 | 29 |
| | Statistics and Probability | – | – | – | 62 | 104 | 77 | 9 |
| | The Real and Complex Number Systems | – | – | – | 191 | 149 | 39 | 4 |
| | **Total** | **498** | **491** | **498** | **435** | **433** | **448** | **63** |

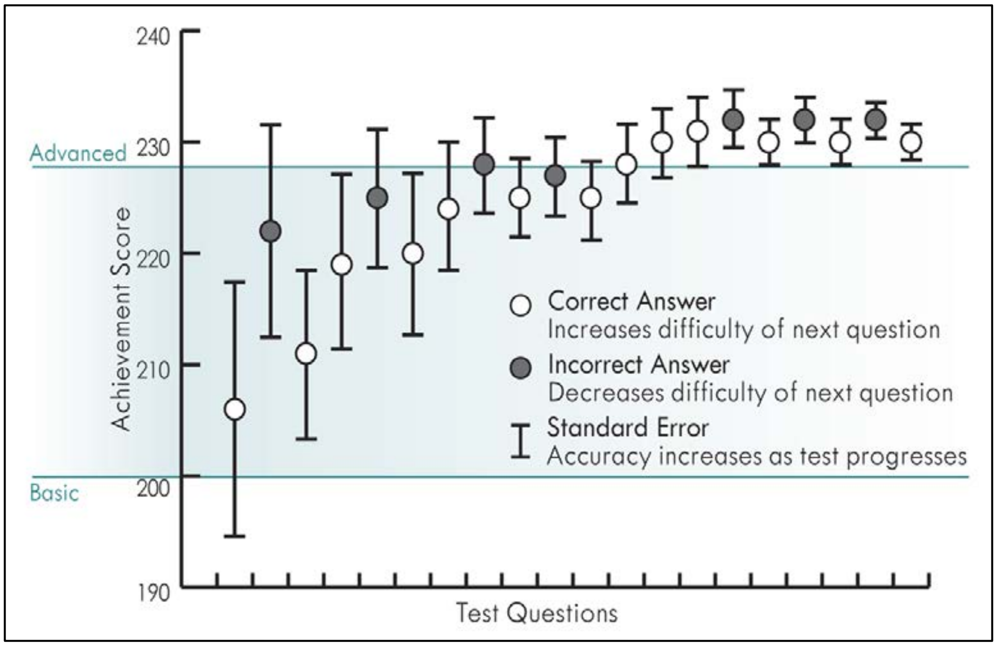
Beyond the instructional areas, the lower standard levels were also examined by assessing the number of items available at each standard. The percentages of students who received at least one item from each standard are shown in Appendix B.

### 3.3. Constraint-Based Adaptive Test Engine

A CAT administers items to match the ability level of the students: different 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.

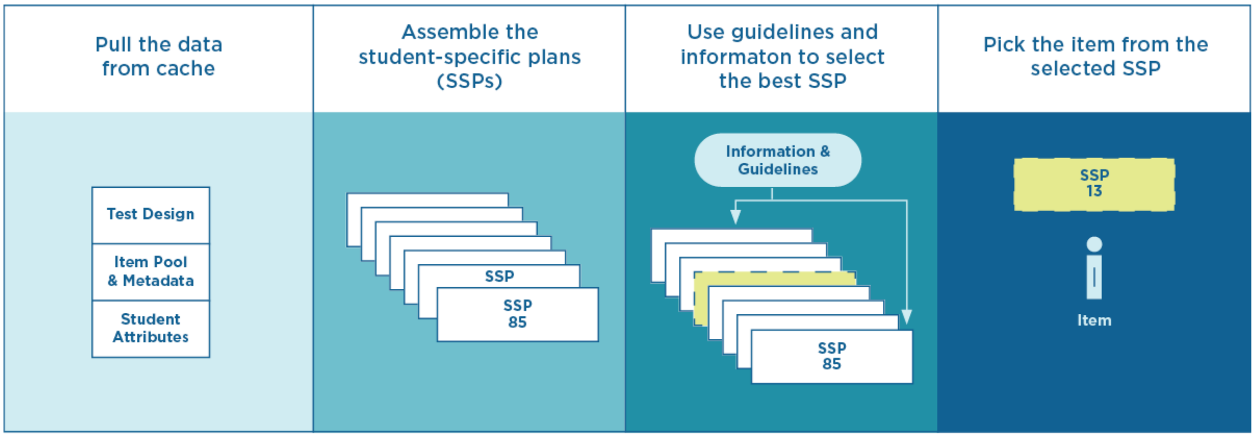
The constraint-based engine (CBE) uses the blueprint and a student's momentary theta ($\theta$) to drive item selection, as shown in Figure 3.1. Momentary theta is the ability estimate of the student that is recalculated and updated after answering each item. Items are selected based on item difficulty. 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**


Pull the data from cache
Assemble the student-specific plans (SSPs)
Use guidelines and informaton to select the best SSP
Pick the item from the selected SSP
Test Design
Item Pool & Metadata
Student Attributes
SSP
SSP 85
Information & Guidelines
SSP 85
SSP 13
Item

*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 a 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 are 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 uses 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 closes, a post-administration evaluation study is 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 rates 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 3.6 presents the blueprint constraint results at the reporting category level for 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 3.6. 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 | 13 | 12 | 14 | 100 |
| | Informational Text | 8 | 9 | 8 | 8 | 9 | 100 |
| | Vocabulary | 5 | 7 | 6 | 5 | 7 | 100 |
| 4 | Literary Text | 11 | 12 | 11 | 11 | 12 | 100 |
| | Informational Text | 9 | 11 | 9 | 9 | 11 | 100 |
| | Vocabulary | 5 | 7 | 7 | 5 | 7 | 100 |
| 5 | Literary Text | 9 | 11 | 10 | 9 | 11 | 100 |
| | Informational Text | 9 | 11 | 10 | 9 | 11 | 100 |
| | Vocabulary | 5 | 7 | 6 | 5 | 7 | 100 |
| 6 | Literary Text | 9 | 11 | 10 | 9 | 11 | 100 |
| | Informational Text | 11 | 12 | 11 | 11 | 12 | 100 |
| | Vocabulary | 5 | 7 | 6 | 5 | 7 | 100 |
| 7 | Literary Text | 8 | 9 | 9 | 8 | 9 | 100 |
| | Informational Text | 12 | 14 | 12 | 12 | 14 | 100 |
| | Vocabulary | 5 | 7 | 6 | 5 | 7 | 100 |
| 8 | Literary Text | 8 | 9 | 8 | 8 | 9 | 100 |
| | Informational Text | 12 | 14 | 12 | 12 | 14 | 100 |

| Grade | Summative Content Across Instructional Areas | #Items Intended | | #Items Administered | | | %Match |
|---|---|---|---|---|---|---|---|
| | | Min | Max | Average | Min | Max | |
| | Vocabulary | 5 | 7 | 7 | 5 | 7 | 100 |
| HS | Literary Text | 8 | 9 | 8 | 8 | 8 | 100 |
| | Informational Text | 12 | 14 | 14 | 14 | 14 | 100 |
| | Vocabulary | 5 | 8 | 8 | 8 | 8 | 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 | 5 | 5 | 100 |
| | Geometry | 4 | 4 | 4 | 4 | 4 | 100 |
| 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 |

| Grade | Summative Content Across Instructional Areas | #Items Intended | | #Items Administered | | | %Match |
|---|---|---|---|---|---|---|---|
| | | Min | Max | Average | Min | Max | |
| | 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 (θ)* is the IRT CSEM for a scale score, and *I(θ)* 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 3.7 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 3.7. CSEMs at the Cut Scores**

| Content Area | Grade | *Below State Expectations* | | | *At State Expectations* | | | *Above State Expectations* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Mean CSEM | SD | N | Mean CSEM | SD | N | Mean CSEM | SD |
| Reading | 3 | 2,864 | 5.13 | 0.37 | 5,382 | 4.69 | 0.53 | 2,670 | 5.02 | 0.15 |
| | 4 | 3,354 | 5.05 | 0.60 | 5,503 | 4.72 | 0.52 | 3,358 | 4.87 | 0.64 |
| | 5 | 2,887 | 5.02 | 0.28 | 4,560 | 4.95 | 0.26 | 3,617 | 4.78 | 0.69 |
| | 6 | 2,543 | 5.09 | 0.47 | 5,044 | 4.81 | 0.39 | 3,379 | 4.84 | 0.45 |
| | 7 | 2,549 | 5.13 | 0.56 | 4,992 | 5.01 | 0.15 | 3,215 | 5.02 | 0.33 |
| | 8 | 2,836 | 4.97 | 0.31 | 5,389 | 4.84 | 0.45 | 3,402 | 4.95 | 0.24 |
| | HS | 4,488 | 5.40 | 0.57 | 5,692 | 5.03 | 0.36 | 2,817 | 6.10 | 0.70 |
| Mathematics | 3 | 2,830 | 5.00 | 0.08 | 5,081 | 4.91 | 0.30 | 2,692 | 4.99 | 0.08 |
| | 4 | 4,372 | 3.82 | 0.40 | 5,736 | 3.78 | 0.43 | 2,428 | 4.01 | 0.14 |
| | 5 | 3,753 | 4.01 | 0.21 | 5,825 | 3.98 | 0.21 | 2,388 | 4.01 | 0.12 |
| | 6 | 3,679 | 5.01 | 0.48 | 5,212 | 4.93 | 0.37 | 1,619 | 4.92 | 0.40 |
| | 7 | 4,514 | 4.01 | 0.41 | 5,167 | 3.99 | 0.12 | 1,630 | 4.00 | 0.76 |

| Content Area | Grade | *Below State Expectations* | | | *At State Expectations* | | | *Above State Expectations* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Mean CSEM | SD | N | Mean CSEM | SD | N | Mean CSEM | SD |
| | 8 | 5,348 | 4.01 | 0.18 | 5,230 | 3.98 | 0.24 | 1,463 | 4.00 | 0.04 |
| | HS | 6,487 | 7.22 | 0.66 | 5,669 | 7.02 | 0.38 | 1,275 | 7.00 | 0.62 |

Table 3.8 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 3.8. CSEMs by Score Decile**

| Grade | Overall | Proficiency Score Decile | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
| **Reading** | | | | | | | | | | | |
| 3 | 4.94 | 5.55 | 5.02 | 5.00 | 4.94 | 4.51 | 4.38 | 4.73 | 5.01 | 5.00 | 5.22 |
| 4 | 4.87 | 5.44 | 5.01 | 5.01 | 4.87 | 4.55 | 4.43 | 4.46 | 4.62 | 4.99 | 5.30 |
| 5 | 4.95 | 5.55 | 5.01 | 5.00 | 5.00 | 4.89 | 4.75 | 4.64 | 4.60 | 4.86 | 5.20 |
| 6 | 4.91 | 5.54 | 5.00 | 5.00 | 4.84 | 4.62 | 4.55 | 4.57 | 4.67 | 4.89 | 5.40 |
| 7 | 5.12 | 5.44 | 5.01 | 5.01 | 5.00 | 5.01 | 5.01 | 5.00 | 5.03 | 5.01 | 5.66 |
| 8 | 4.92 | 5.17 | 5.02 | 4.85 | 4.87 | 4.90 | 4.73 | 4.63 | 4.85 | 5.00 | 5.20 |
| HS | 5.53 | 6.25 | 5.71 | 5.02 | 5.04 | 5.02 | 5.03 | 5.01 | 5.45 | 6.01 | 6.79 |
| **Mathematics** | | | | | | | | | | | |
| 3 | 4.98 | 5.22 | 5.00 | 5.00 | 4.99 | 4.83 | 4.84 | 4.90 | 4.98 | 5.00 | 5.07 |
| 4 | 3.91 | 4.06 | 4.00 | 3.81 | 3.67 | 3.64 | 3.85 | 3.99 | 3.99 | 4.01 | 4.05 |
| 5 | 4.01 | 4.06 | 4.01 | 4.01 | 4.00 | 3.98 | 3.93 | 3.99 | 4.00 | 4.01 | 4.06 |
| 6 | 4.96 | 5.12 | 5.02 | 5.01 | 5.00 | 5.00 | 4.94 | 4.98 | 4.77 | 4.69 | 5.03 |
| 7 | 4.03 | 4.23 | 4.04 | 4.01 | 4.00 | 4.00 | 4.00 | 4.00 | 3.96 | 3.97 | 4.04 |
| 8 | 4.03 | 4.28 | 4.01 | 4.01 | 4.00 | 4.01 | 4.01 | 3.94 | 3.96 | 4.00 | 4.04 |
| HS | 7.28 | 9.35 | 7.17 | 7.01 | 7.02 | 7.02 | 7.01 | 7.02 | 7.03 | 7.02 | 7.14 |

### *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. The exposure rate for each operational item was calculated as the percentage of students who received that item, as shown in Table 3.9. 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.

**Table 3.9. Operational Item Exposure Rates**

| Grade | #Items | | | | Item Exposure Rate | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 0–20% | | 21–40% | | 41–60% | | 61–80% | | 81–99% | | 100% | |
| | Total | Used | Unused | Unused % | N | % | N | % | N | % | N | % | N | % | N | % |
| **Reading** | | | | | | | | | | | | | | | | |
| 3 | 661 | 383 | 278 | 42.06 | 331 | 86.42 | 37 | 9.66 | 11 | 2.87 | 2 | 0.52 | 0 | 0 | 0 | 0 |
| 4 | 985 | 493 | 492 | 49.95 | 459 | 93.10 | 25 | 5.07 | 4 | 0.81 | 2 | 0.41 | 0 | 0 | 0 | 0 |
| 5 | 898 | 497 | 401 | 44.65 | 464 | 93.36 | 21 | 4.23 | 11 | 2.21 | 1 | 0.20 | 0 | 0 | 0 | 0 |
| 6 | 890 | 450 | 440 | 49.44 | 421 | 93.56 | 11 | 2.44 | 6 | 1.33 | 8 | 1.78 | 0 | 0 | 0 | 0 |
| 7 | 961 | 499 | 462 | 48.07 | 470 | 94.19 | 20 | 4.01 | 3 | 0.60 | 2 | 0.40 | 0 | 0 | 0 | 0 |
| 8 | 686 | 464 | 222 | 32.36 | 426 | 91.81 | 34 | 7.33 | 4 | 0.86 | 0 | 0.00 | 0 | 0 | 0 | 0 |
| HS | 49 | 49 | 0 | 0.00 | 5 | 10.20 | 9 | 18.37 | 12 | 24.49 | 5 | 10.20 | 0 | 0 | 0 | 0 |
| **Mathematics** | | | | | | | | | | | | | | | | |
| 3 | 742 | 498 | 244 | 32.88 | 483 | 96.99 | 12 | 2.41 | 0 | 0.00 | 1 | 0.20 | 0 | 0 | 0 | 0 |
| 4 | 1046 | 491 | 555 | 53.06 | 479 | 97.56 | 12 | 2.44 | 0 | 0.00 | 0 | 0.00 | 0 | 0 | 0 | 0 |
| 5 | 962 | 498 | 464 | 48.23 | 485 | 97.39 | 10 | 2.01 | 3 | 0.60 | 0 | 0.00 | 0 | 0 | 0 | 0 |
| 6 | 949 | 435 | 514 | 54.16 | 418 | 96.09 | 16 | 3.68 | 1 | 0.23 | 0 | 0.00 | 0 | 0 | 0 | 0 |
| 7 | 956 | 433 | 523 | 54.71 | 410 | 94.69 | 21 | 4.85 | 1 | 0.23 | 1 | 0.23 | 0 | 0 | 0 | 0 |
| 8 | 687 | 448 | 239 | 34.79 | 432 | 96.43 | 11 | 2.46 | 4 | 0.89 | 1 | 0.22 | 0 | 0 | 0 | 0 |
| HS | 63 | 63 | 0 | 0.00 | 29 | 46.03 | 4 | 6.35 | 0 | 0.00 | 7 | 11.11 | 0 | 0 | 0 | 0 |

A number of field test items were embedded in the Spring 2024 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 3.10 show that all field test items were appropriately administered to each demographic subgroup.

**Table 3.10. Field Test Item Exposure Rates**

| Grade | FT Items | Mean | Female | Male | Hispanic/ Latino | Am. Indian or Alaska Native | Asian | Black or African American | Native HI or Pacific Islander | White | Two or More Races |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reading** | | | | | | | | | | | |
| 3 | 5 | 11,651 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 4 | 5 | 12,210 | 49 | 51 | 4 | 1 | 1 | 5 | 0 | 86 | 4 |
| 5 | 7 | 8,707 | 48 | 52 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 6 | 5 | 11,943 | 49 | 51 | 3 | 1 | 1 | 4 | 0 | 86 | 4 |
| 7 | 5 | 12,190 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 8 | 12 | 5,116 | 48 | 52 | 3 | 1 | 1 | 5 | 0 | 86 | 3 |
| HS | 117 | 747 | 48 | 52 | 3 | 1 | 2 | 5 | 0 | 86 | 3 |
| **Mathematics** | | | | | | | | | | | |
| 3 | 6 | 9,766 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 85 | 4 |
| 4 | 11 | 5,582 | 49 | 51 | 4 | 1 | 1 | 5 | 0 | 86 | 4 |
| 5 | 5 | 12,250 | 48 | 52 | 3 | 1 | 1 | 5 | 0 | 85 | 4 |
| 6 | 18 | 3,329 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 7 | 19 | 3,217 | 49 | 51 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| 8 | 13 | 4,740 | 48 | 52 | 3 | 1 | 1 | 5 | 0 | 86 | 4 |
| HS | 145 | 603 | 48 | 52 | 3 | 1 | 2 | 5 | 0 | 86 | 3 |

*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, 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 Spring 2024, 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 and large print forms were prepared in advance according to registration data, and the required materials were packed and shipped to the requesting schools. Standard paper-based forms were available via print on demand. These materials were sent to School Assessment Coordinators via NWEA's secure SFTP site.

Table 3.11 presents the numbers 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 3.11. Paper Form Summative Item Totals by Content and Grade**

| Content | Grade | Summative Operational |
|---|---|---|
| Reading | 3−8 | 27 |
| Math | 3−8 | 27 |
| Reading | HS | 30 |
| Math | HS | 30 |

### 3.5. Spring 2024 Fixed Form Blueprints

The blueprints of the fixed forms are consistent with those of the online adaptive forms in terms of the item count per instructional area as shown in Section 2.2 and Appendix G. Table 3.12–Table 3.16 display the item counts by instructional area for the Spring 2024 assessments. Please note that small discrepancies may appear between the data in these tables and that in Section 2.2 and Appendix G due to rounding and the nature of approximations for ranges.

**Table 3.12. Reading Item Counts by Instructional Area, Grades 3–8**

| **Instructional Area** | **Total** | | | | | | **Summative** | | | | | | **Diagnostic** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **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 (44%) | 11 (41%) | 11 (41%) | 9 (33%) | 8 (30%) | 8 (30%) | 4 | 4 | 4 | 5 | 5 | 5 |
| Informational Text | 13 | 14 | 14 | 15 | 16 | 16 | 8 (30%) | 9 (33%) | 9 (33%) | 11 (41%) | 12 (44%) | 12 (44%) | 5 | 5 | 5 | 4 | 4 | 4 |
| Vocabulary | 12 | 12 | 12 | 12 | 12 | 12 | 7 (26%) | 7 (26%) | 7 (26%) | 7 (26%) | 7 (26%) | 7 (26%) | 5 | 5 | 5 | 5 | 5 | 5 |
| **Total Items** | **41** | **41** | **41** | **41** | **41** | **41** | **27 (100%)** | **27 (100%)** | **27 (100%)** | **27 (100%)** | **27 (100%)** | **27 (100%)** | **14** | **14** | **14** | **14** | **14** | **14** |

**Table 3.13. Reading Item Counts by Instructional Area, HS**

| Instructional Area | Total | Summative | Diagnostic |
|---|---|---|---|
| | HS | HS | HS |
| Literary Text | 13 | 9<br>(30%) | 4 |
| Informational Text | 17 | 13<br>(43%) | 4 |
| Vocabulary | 12 | 8<br>(27%) | 4 |
| **Total Items** | **42** | **30<br>(100%)** | **12** |

**Table 3.14. 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<br>(22%) | 5<br>(18.5%) | 4<br>(15%) | 4 | 5 | 5 |
| Numbers and Operations | 13 | 17 | 18 | 9<br>(33%) | 13<br>(48%) | 14<br>(52%) | 4 | 4 | 4 |
| Measurement and Data | 12 | 9 | 9 | 8<br>(30%) | 5<br>(18.5%) | 5<br>(18%) | 4 | 4 | 4 |
| Geometry | 10 | 9 | 9 | 4<br>(15%) | 4<br>(15%) | 4<br>(15%) | 6 | 5 | 5 |
| **Total Items** | **45** | **45** | **45** | **27<br>(100%)** | **27<br>(100%)** | **27<br>(100%)** | **18** | **18** | **18** |

**Table 3.15. 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<br>(26%) | 5<br>(18.5%) | 13<br>(48%) | 4 | 5 | 4 |
| The Real and Complex Number Systems | 16 | 15 | 9 | 12<br>(44%) | 11<br>(41%) | 4<br>(15%) | 4 | 4 | 5 |
| Geometry | 9 | 10 | 10 | 4<br>(15%) | 6<br>(22%) | 6<br>(22%) | 5 | 4 | 4 |
| Statistics and Probability | 9 | 10 | 9 | 4<br>(15%) | 5<br>(18.5%) | 4<br>(15%) | 5 | 5 | 5 |
| **Total Items** | **45** | **45** | **45** | **27<br>(100%)** | **27<br>(100%)** | **27<br>(100%)** | **18** | **18** | **18** |

**Table 3.16. Mathematics Item Counts by Instructional Area, HS**

| Instructional Area | Total | Summative | Diagnostic |
|---|---|---|---|
| | HS | HS | HS |
| Operations and Algebraic Thinking | 17 | 13<br>(43%) | 4 |
| The Real and Complex Number Systems | 9 | 5<br>(17%) | 4 |
| Geometry | 12 | 8<br>(27%) | 4 |
| Statistics and Probability | 9 | 4<br>(13%) | 5 |
| **Total Items** | **47** | **30<br>(100%)** | **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 or large print forms immediately upon receipt to inventory the contents. School Assessment Coordinators are responsible for the printing and secure handling of standard 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 are 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 periodically to gain insights into the information contained within them.
- 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 regularly for any unexpected events, including hacking attempts and attacks.

# Section 4: Item Statistics, Calibration, and Scaling

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, respectively, 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 all grades. 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 | 192 | 0.47 | 0.46 | 0.12 | 0.12 | 0.88 | 0 | 1 | 12 | 48 | 54 | 47 | 24 | 3 | 3 | 0 |
| 4 | 256 | 0.47 | 0.45 | 0.13 | 0.12 | 0.93 | 0 | 3 | 18 | 53 | 90 | 56 | 22 | 9 | 4 | 1 |
| 5 | 272 | 0.48 | 0.47 | 0.13 | 0.05 | 0.86 | 1 | 2 | 12 | 47 | 105 | 62 | 26 | 13 | 4 | 0 |
| 6 | 221 | 0.48 | 0.47 | 0.12 | 0.19 | 0.86 | 0 | 1 | 6 | 45 | 83 | 52 | 24 | 7 | 3 | 0 |
| 7 | 251 | 0.46 | 0.45 | 0.12 | 0.13 | 0.90 | 0 | 2 | 18 | 48 | 102 | 52 | 20 | 7 | 1 | 1 |
| 8 | 306 | 0.50 | 0.51 | 0.13 | 0.10 | 0.95 | 1 | 3 | 12 | 39 | 94 | 92 | 53 | 8 | 3 | 1 |
| HS | 48 | 0.50 | 0.47 | 0.13 | 0.19 | 0.84 | 0 | 1 | 1 | 8 | 19 | 6 | 10 | 2 | 1 | 0 |
| **Mathematics** | | | | | | | | | | | | | | | | |
| 3 | 422 | 0.48 | 0.48 | 0.10 | 0.10 | 0.78 | 1 | 3 | 15 | 59 | 180 | 124 | 32 | 8 | 0 | 0 |
| 4 | 391 | 0.47 | 0.47 | 0.11 | 0.05 | 0.91 | 1 | 3 | 14 | 69 | 157 | 106 | 29 | 11 | 0 | 1 |
| 5 | 408 | 0.46 | 0.46 | 0.11 | 0.02 | 0.84 | 3 | 5 | 15 | 81 | 166 | 110 | 21 | 6 | 1 | 0 |
| 6 | 357 | 0.46 | 0.46 | 0.10 | 0.14 | 0.81 | 0 | 2 | 17 | 76 | 148 | 89 | 18 | 5 | 2 | 0 |
| 7 | 349 | 0.46 | 0.46 | 0.09 | 0.13 | 0.71 | 0 | 4 | 12 | 66 | 163 | 89 | 13 | 2 | 0 | 0 |
| 8 | 392 | 0.46 | 0.46 | 0.09 | 0.14 | 0.74 | 0 | 3 | 19 | 70 | 191 | 90 | 18 | 1 | 0 | 0 |
| HS | 63 | 0.38 | 0.35 | 0.15 | 0.04 | 0.74 | 2 | 5 | 14 | 20 | 7 | 9 | 4 | 2 | 0 | 0 |

**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 | 5 | 0.34 | 0.34 | 0.10 | 0.24 | 0.50 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| 4 | 5 | 0.36 | 0.37 | 0.08 | 0.25 | 0.45 | 0 | 0 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 7 | 0.45 | 0.48 | 0.20 | 0.19 | 0.78 | 0 | 1 | 0 | 2 | 2 | 1 | 0 | 1 | 0 | 0 |
| 6 | 5 | 0.44 | 0.46 | 0.06 | 0.34 | 0.49 | 0 | 0 | 0 | 1 | 4 | 0 | 0 | 0 | 0 | 0 |
| 7 | 5 | 0.47 | 0.50 | 0.14 | 0.23 | 0.62 | 0 | 0 | 1 | 0 | 2 | 1 | 1 | 0 | 0 | 0 |
| 8 | 12 | 0.44 | 0.42 | 0.17 | 0.21 | 0.83 | 0 | 0 | 2 | 3 | 4 | 1 | 1 | 0 | 1 | 0 |
| HS | 117 | 0.47 | 0.47 | 0.13 | 0.13 | 0.81 | 0 | 4 | 9 | 19 | 37 | 28 | 17 | 2 | 1 | 0 |
| **Mathematics** | | | | | | | | | | | | | | | | |
| 3 | 6 | 0.24 | 0.21 | 0.12 | 0.12 | 0.47 | 0 | 2 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 4 | 11 | 0.32 | 0.28 | 0.16 | 0.12 | 0.60 | 0 | 3 | 3 | 2 | 0 | 3 | 0 | 0 | 0 | 0 |
| 5 | 5 | 0.24 | 0.27 | 0.12 | 0.06 | 0.38 | 1 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 18 | 0.16 | 0.12 | 0.14 | 0.01 | 0.46 | 9 | 4 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| 7 | 19 | 0.13 | 0.09 | 0.11 | 0.03 | 0.49 | 10 | 6 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 8 | 13 | 0.18 | 0.16 | 0.14 | 0.04 | 0.57 | 3 | 7 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| HS | 145 | 0.23 | 0.20 | 0.15 | 0.00 | 0.78 | 34 | 39 | 30 | 23 | 10 | 6 | 2 | 1 | 0 | 0 |

### *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 | 192 | 0.35 | 0.35 | 0.12 | 0.00 | 0.69 | 6 | 13 | 38 | 73 | 42 | 17 | 3 | 0 | 3 | 1 |
| 4 | 256 | 0.35 | 0.35 | 0.13 | -0.16 | 0.68 | 9 | 17 | 54 | 94 | 57 | 21 | 4 | 0 | 5 | 5 |
| 5 | 272 | 0.35 | 0.35 | 0.12 | -0.37 | 0.70 | 5 | 19 | 50 | 118 | 55 | 19 | 5 | 1 | 0 | 2 |
| 6 | 221 | 0.36 | 0.36 | 0.10 | 0.00 | 0.71 | 2 | 9 | 51 | 85 | 54 | 19 | 0 | 1 | 0 | 6 |
| 7 | 251 | 0.34 | 0.34 | 0.12 | 0.00 | 0.66 | 7 | 18 | 61 | 91 | 56 | 12 | 6 | 0 | 4 | 5 |
| 8 | 306 | 0.36 | 0.36 | 0.11 | -0.07 | 0.64 | 7 | 18 | 56 | 119 | 73 | 30 | 3 | 0 | 3 | 5 |
| HS | 48 | 0.39 | 0.38 | 0.14 | -0.07 | 0.65 | 1 | 1 | 13 | 10 | 11 | 10 | 2 | 0 | 0 | 0 |
| **Mathematics** | | | | | | | | | | | | | | | | |
| 3 | 422 | 0.35 | 0.35 | 0.08 | -0.09 | 0.66 | 3 | 4 | 93 | 217 | 89 | 12 | 4 | 0 | 0 | 0 |
| 4 | 391 | 0.36 | 0.36 | 0.09 | -0.08 | 0.75 | 5 | 8 | 77 | 196 | 82 | 21 | 1 | 1 | 0 | 0 |
| 5 | 408 | 0.35 | 0.36 | 0.09 | -0.16 | 0.65 | 4 | 13 | 90 | 188 | 89 | 19 | 5 | 0 | 0 | 0 |
| 6 | 357 | 0.36 | 0.36 | 0.10 | -0.01 | 0.71 | 2 | 10 | 77 | 151 | 91 | 22 | 3 | 1 | 0 | 0 |
| 7 | 349 | 0.35 | 0.36 | 0.09 | -0.03 | 0.63 | 2 | 16 | 67 | 166 | 82 | 15 | 1 | 0 | 3 | 0 |
| 8 | 392 | 0.34 | 0.34 | 0.09 | -0.09 | 0.80 | 2 | 10 | 111 | 192 | 61 | 13 | 2 | 0 | 1 | 0 |
| HS | 63 | 0.37 | 0.39 | 0.12 | 0.07 | 0.66 | 2 | 2 | 13 | 19 | 18 | 6 | 3 | 0 | 0 | 0 |

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 | 5 | 0.26 | 0.22 | 0.07 | 0.21 | 0.35 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 5 | 0.23 | 0.19 | 0.08 | 0.15 | 0.34 | 0 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 7 | 0.29 | 0.37 | 0.18 | 0.05 | 0.47 | 2 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| 6 | 5 | 0.27 | 0.25 | 0.09 | 0.16 | 0.37 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 5 | 0.28 | 0.28 | 0.03 | 0.26 | 0.33 | 0 | 0 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 12 | 0.23 | 0.22 | 0.11 | 0.08 | 0.45 | 2 | 3 | 3 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| HS | 117 | 0.35 | 0.37 | 0.13 | 0.01 | 0.66 | 6 | 8 | 19 | 39 | 34 | 10 | 1 | 0 | 0 | 0 |
| **Mathematics** | | | | | | | | | | | | | | | | |
| 3 | 6 | 0.16 | 0.16 | 0.06 | 0.08 | 0.22 | 1 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 11 | 0.20 | 0.17 | 0.13 | -0.02 | 0.42 | 2 | 4 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 0.20 | 0.16 | 0.13 | 0.09 | 0.39 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 18 | 0.21 | 0.24 | 0.14 | -0.05 | 0.49 | 5 | 2 | 8 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| 7 | 19 | 0.26 | 0.25 | 0.15 | -0.01 | 0.54 | 4 | 3 | 3 | 4 | 4 | 1 | 0 | 0 | 0 | 0 |
| 8 | 13 | 0.33 | 0.35 | 0.15 | 0.06 | 0.58 | 1 | 3 | 1 | 4 | 3 | 1 | 0 | 0 | 0 | 0 |
| HS | 145 | 0.35 | 0.37 | 0.15 | -0.02 | 0.66 | 7 | 19 | 22 | 32 | 45 | 15 | 5 | 0 | 0 | 0 |

## 4.2. IRT Calibration

When Maine’s scale was established in 2023, the first step was to calibrate items to a standardized scale and 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 $\theta$ responding correctly to item $i$ is as follows, where $\theta_j$ and $b_i$ are the person and item parameters, respectively:

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

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

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

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 $\theta_j$ and $b_i$ are the person and item parameters, respectively.

The free calibration[1] 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 |

[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) |
|---|---|---|---|---|---|---|---|---|
| 6 | 682 | 769 | 0.23 | 0.21 | 1.79 | -4.57 | 6.37 | 10.94 |
| 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 |

### 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 cluing.

#### *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 review process pays more attention to the infit than to the outfit because infit is the more stable statistic.

The table shows that all average infit and outfit values are close to 1.0, indicating that items fit well at their intended grade level. Infits are very stable across grades, as they reflect how well an individual item fits the overall measurement model. This stability ensures that results are not significantly affected by grade-level differences, making it particularly useful for longitudinal or multi-grade assessments. 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 |

*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 Commissioner of Education.
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. It is worth noting that only summative items are included in the calculation of state summative scale scores. The summative item counts are 27 for grades 3–8 and 30 for the second year of high school, as presented in Table 3.6. The pattern scoring of a CAT produces many more scale score points, while a fixed form assessment has a finite number of scale scores corresponding to the raw scores.

**Table 4.7. Maine Grade-Level Scale Properties**

| Grade | Scaling Constants | | Scale Score Cuts | | | Range | |
|---|---|---|---|---|---|---|---|
| | Intercept | Slope | *Below State Expectations* | *At State Expectations* | *Above State Expectations* | 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 |

| Grade | Scaling Constants | | Scale Score Cuts | | | Range | |
|---|---|---|---|---|---|---|---|
| | Intercept | Slope | *Below State Expectations* | *At State Expectations* | *Above State Expectations* | LOSS | HOSS |
| 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: 2, 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, 2015) 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, 5, 7, 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 mathematics.
- 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 test results.
- Relevant sections within this report: 2, 7

**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, 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.
- 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: 2, 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 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, 7, 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 Content 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<br>4. Item Statistics, Calibration, and Scaling<br>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<br>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 Content Development<br>3. Administration and Security<br>6. Technical Quality—Other<br>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, p. 1-16).

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. Samejima (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:

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

where σ is defined as:

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

Table 6.1 and Table 6.2 present the overall error of estimated theta and test reliability for the grades 3–8 and second year of high school 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 0.87 or higher for reading and 0.83 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 | 1.34 | 0.39 | 0.91 |
| 4 | 27 | 1.27 | 0.39 | 0.91 |
| 5 | 27 | 1.23 | 0.39 | 0.90 |
| 6 | 27 | 1.13 | 0.39 | 0.88 |
| 7 | 27 | 1.22 | 0.41 | 0.89 |
| 8 | 27 | 1.34 | 0.39 | 0.91 |
| HS | 30 | 0.97 | 0.35 | 0.87 |

**Table 6.2. Reliability Statistics—Mathematics**

| Grade | Average # Items | SD of Estimated Theta | Mean SEM | Reliability |
|---|---|---|---|---|
| 3 | 27 | 1.59 | 0.40 | 0.94 |
| 4 | 27 | 1.72 | 0.39 | 0.95 |
| 5 | 27 | 1.65 | 0.40 | 0.94 |
| 6 | 27 | 1.64 | 0.40 | 0.94 |
| 7 | 27 | 1.69 | 0.40 | 0.94 |
| 8 | 27 | 1.50 | 0.40 | 0.93 |
| HS | 30 | 1.02 | 0.41 | 0.83 |

### *6.1.2. 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 6.3 and Table 6.4 present the respective reading and mathematics classification accuracy results by grade and achievement level for grades 3–8 and the second year of high school. Overall classification accuracy ranges from 0.69 to 0.87.

Table 6.5 presents the classification *accuracy* results by grade at each achievement level and each cut for grades 3–8 and the second year of high school. Overall classification accuracy ranges from 0.69 to 0.87. The classification accuracy at each achievement level ranges from 0.52 to 0.93, whereas the classification accuracy at each cut ranges from 0.85 to 0.98.

Table 6.6 presents the classification *consistency* results by grade at each achievement level and each cut for grades 3–8 and the second year of high school. Overall classification consistency ranges from 0.60 to 0.81. The classification consistency at each achievement level ranges from 0.43 to 0.85, whereas the classification consistency at each cut ranges from 0.80 to 0.97.

Regarding the lower levels of classification accuracy and consistency for the second year of high school tests, NWEA is conducting investigations on improving measurement precision to be implemented in 2025–2026.

**Table 6.3. Classification Accuracy by Achievement Level—Reading**

| Grade | Achievement Level | N | Prop. | Expected Proportion [a] | | | | Class. Acc. | Overall Class. Acc. |
|---|---|---|---|---|---|---|---|---|---|
| | | | | L1 | L2 | L3 | L4 | | |
| 3 | *Well-Below State Expectations* | 1,187 | 0.10 | 0.08 | 0.02 | 0.00 | 0.00 | 0.82 | 0.81 |
| | *Below State Expectations* | 3,477 | 0.30 | 0.03 | 0.23 | 0.04 | 0.00 | 0.78 | |
| | *At State Expectations* | 5,703 | 0.49 | 0.00 | 0.05 | 0.41 | 0.02 | 0.84 | |
| | *Above State Expectations* | 1,300 | 0.11 | 0.00 | 0.00 | 0.02 | 0.09 | 0.79 | |
| 4 | *Well-Below State Expectations* | 1,449 | 0.12 | 0.10 | 0.02 | 0.00 | 0.00 | 0.83 | 0.81 |
| | *Below State Expectations* | 2,946 | 0.24 | 0.03 | 0.17 | 0.04 | 0.00 | 0.72 | |
| | *At State Expectations* | 6,468 | 0.53 | 0.00 | 0.05 | 0.45 | 0.03 | 0.86 | |
| | *Above State Expectations* | 1,373 | 0.11 | 0.00 | 0.00 | 0.02 | 0.09 | 0.79 | |
| 5 | *Well-Below State Expectations* | 1,403 | 0.11 | 0.10 | 0.01 | 0.00 | 0.00 | 0.87 | 0.82 |
| | *Below State Expectations* | 2,258 | 0.18 | 0.03 | 0.13 | 0.03 | 0.00 | 0.69 | |
| | *At State Expectations* | 7,009 | 0.57 | 0.00 | 0.05 | 0.50 | 0.03 | 0.86 | |
| | *Above State Expectations* | 1,543 | 0.13 | 0.00 | 0.00 | 0.03 | 0.10 | 0.80 | |
| 6 | *Well-Below State Expectations* | 928 | 0.08 | 0.06 | 0.01 | 0.00 | 0.00 | 0.83 | 0.82 |
| | *Below State Expectations* | 2446 | 0.20 | 0.03 | 0.15 | 0.03 | 0.00 | 0.71 | |
| | *At State Expectations* | 7168 | 0.60 | 0.00 | 0.05 | 0.52 | 0.03 | 0.86 | |
| | *Above State Expectations* | 1415 | 0.12 | 0.00 | 0.00 | 0.03 | 0.09 | 0.78 | |
| 7 | *Well-Below State Expectations* | 792 | 0.06 | 0.05 | 0.01 | 0.00 | 0.00 | 0.79 | 0.82 |
| | *Below State Expectations* | 3,203 | 0.26 | 0.03 | 0.20 | 0.03 | 0.00 | 0.77 | |
| | *At State Expectations* | 6,539 | 0.54 | 0.00 | 0.05 | 0.46 | 0.03 | 0.85 | |
| | *Above State Expectations* | 1,676 | 0.14 | 0.00 | 0.00 | 0.02 | 0.11 | 0.82 | |
| 8 | *Well-Below State Expectations* | 1,396 | 0.11 | 0.10 | 0.01 | 0.00 | 0.00 | 0.88 | 0.82 |
| | *Below State Expectations* | 3,100 | 0.25 | 0.03 | 0.19 | 0.03 | 0.00 | 0.76 | |
| | *At State Expectations* | 6,258 | 0.51 | 0.00 | 0.05 | 0.43 | 0.03 | 0.85 | |
| | *Above State Expectations* | 1,540 | 0.13 | 0.00 | 0.00 | 0.03 | 0.10 | 0.79 | |
| HS | *Well-Below State Expectations* | 2,113 | 0.17 | 0.14 | 0.03 | 0.00 | 0.00 | 0.83 | 0.78 |
| | *Below State Expectations* | 2,806 | 0.22 | 0.05 | 0.14 | 0.03 | 0.00 | 0.64 | |
| | *At State Expectations* | 6,524 | 0.52 | 0.00 | 0.05 | 0.43 | 0.03 | 0.83 | |

| Grade | Achievement Level | N | Prop. | Expected Proportion [a] | | | | Class. Acc. | Overall Class. Acc. |
|---|---|---|---|---|---|---|---|---|---|
| | | | | L1 | L2 | L3 | L4 | | |
| | *Above State Expectations* | 1,069 | 0.09 | 0.00 | 0.00 | 0.02 | 0.07 | 0.77 | |

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

**Table 6.4. Classification Accuracy by Achievement Level—Mathematics**

| Grade | Achievement Level | N | Prop. | Expected Proportion [a] | | | | Class. Acc. | Overall Class. Acc. |
|---|---|---|---|---|---|---|---|---|---|
| | | | | L1 | L2 | L3 | L4 | | |
| 3 | *Well-Below State Expectations* | 1,961 | 0.17 | 0.15 | 0.02 | 0.00 | 0.00 | 0.90 | 0.83 |
| | *Below State Expectations* | 2,443 | 0.21 | 0.03 | 0.15 | 0.03 | 0.00 | 0.72 | |
| | *At State Expectations* | 5,580 | 0.48 | 0.00 | 0.05 | 0.40 | 0.02 | 0.84 | |
| | *Above State Expectations* | 1,736 | 0.15 | 0.00 | 0.00 | 0.02 | 0.13 | 0.86 | |
| 4 | *Well-Below State Expectations* | 2,370 | 0.19 | 0.17 | 0.02 | 0.00 | 0.00 | 0.90 | 0.84 |
| | *Below State Expectations* | 3,221 | 0.26 | 0.03 | 0.20 | 0.03 | 0.00 | 0.76 | |
| | *At State Expectations* | 5,448 | 0.44 | 0.00 | 0.04 | 0.38 | 0.02 | 0.86 | |
| | *Above State Expectations* | 1,251 | 0.10 | 0.00 | 0.00 | 0.01 | 0.09 | 0.87 | |
| 5 | *Well-Below State Expectations* | 2,030 | 0.17 | 0.15 | 0.02 | 0.00 | 0.00 | 0.88 | 0.85 |
| | *Below State Expectations* | 3,823 | 0.31 | 0.03 | 0.25 | 0.04 | 0.00 | 0.80 | |
| | *At State Expectations* | 5,277 | 0.43 | 0.00 | 0.05 | 0.37 | 0.01 | 0.86 | |
| | *Above State Expectations* | 1,130 | 0.09 | 0.00 | 0.00 | 0.01 | 0.08 | 0.87 | |
| 6 | *Well-Below State Expectations* | 2,337 | 0.19 | 0.18 | 0.02 | 0.00 | 0.00 | 0.91 | 0.83 |
| | *Below State Expectations* | 4,553 | 0.38 | 0.04 | 0.30 | 0.04 | 0.00 | 0.79 | |
| | *At State Expectations* | 4,423 | 0.37 | 0.00 | 0.05 | 0.31 | 0.01 | 0.83 | |
| | *Above State Expectations* | 682 | 0.06 | 0.00 | 0.00 | 0.01 | 0.05 | 0.84 | |
| 7 | *Well-Below State Expectations* | 2,589 | 0.21 | 0.20 | 0.01 | 0.00 | 0.00 | 0.93 | 0.87 |
| | *Below State Expectations* | 4,979 | 0.41 | 0.03 | 0.34 | 0.03 | 0.00 | 0.84 | |
| | *At State Expectations* | 3,778 | 0.31 | 0.00 | 0.03 | 0.27 | 0.01 | 0.86 | |
| | *Above State Expectations* | 894 | 0.07 | 0.00 | 0.00 | 0.01 | 0.06 | 0.85 | |
| 8 | *Well-Below State Expectations* | 2,619 | 0.21 | 0.19 | 0.02 | 0.00 | 0.00 | 0.89 | 0.84 |
| | *Below State Expectations* | 4,919 | 0.40 | 0.05 | 0.32 | 0.03 | 0.00 | 0.80 | |

| Grade | Achievement Level | N | Prop. | Expected Proportion [a] | | | | Class. Acc. | Overall Class. Acc. |
|---|---|---|---|---|---|---|---|---|---|
| | | | | L1 | L2 | L3 | L4 | | |
| | *At State Expectations* | 4,145 | 0.34 | 0.00 | 0.04 | 0.29 | 0.01 | 0.86 | |
| | *Above State Expectations* | 653 | 0.05 | 0.00 | 0.00 | 0.01 | 0.04 | 0.83 | |
| HS | *Well-Below State Expectations* | 3,166 | 0.25 | 0.19 | 0.05 | 0.00 | 0.00 | 0.77 | 0.69 |
| | *Below State Expectations* | 4,015 | 0.32 | 0.09 | 0.17 | 0.06 | 0.00 | 0.52 | |
| | *At State Expectations* | 4,067 | 0.32 | 0.01 | 0.06 | 0.24 | 0.01 | 0.75 | |
| | *Above State Expectations* | 1,307 | 0.10 | 0.00 | 0.00 | 0.01 | 0.09 | 0.86 | |

[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 and Cut**

| Grade | Accuracy at AL | | | | Accuracy at Cut | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | AL1 | AL2 | AL3 | AL4 | Cut1 | Cut2 | Cut3 | |
| **Mathematics** | | | | | | | | |
| 3 | 0.90 | 0.72 | 0.84 | 0.86 | 0.96 | 0.91 | 0.96 | 0.83 |
| 4 | 0.90 | 0.76 | 0.86 | 0.87 | 0.95 | 0.93 | 0.97 | 0.84 |
| 5 | 0.88 | 0.80 | 0.86 | 0.87 | 0.95 | 0.92 | 0.97 | 0.85 |
| 6 | 0.91 | 0.79 | 0.83 | 0.84 | 0.94 | 0.91 | 0.98 | 0.83 |
| 7 | 0.93 | 0.84 | 0.86 | 0.85 | 0.95 | 0.93 | 0.98 | 0.87 |
| 8 | 0.89 | 0.80 | 0.86 | 0.83 | 0.92 | 0.93 | 0.98 | 0.84 |
| HS | 0.77 | 0.52 | 0.75 | 0.86 | 0.85 | 0.86 | 0.97 | 0.69 |
| **Reading** | | | | | | | | |
| 3 | 0.82 | 0.78 | 0.84 | 0.79 | 0.95 | 0.91 | 0.95 | 0.81 |
| 4 | 0.83 | 0.72 | 0.86 | 0.79 | 0.95 | 0.92 | 0.95 | 0.81 |
| 5 | 0.87 | 0.69 | 0.86 | 0.80 | 0.95 | 0.92 | 0.94 | 0.82 |
| 6 | 0.83 | 0.71 | 0.86 | 0.78 | 0.96 | 0.92 | 0.94 | 0.82 |
| 7 | 0.79 | 0.77 | 0.85 | 0.82 | 0.96 | 0.92 | 0.95 | 0.82 |
| 8 | 0.88 | 0.76 | 0.85 | 0.79 | 0.96 | 0.92 | 0.95 | 0.82 |
| HS | 0.83 | 0.64 | 0.83 | 0.77 | 0.92 | 0.91 | 0.95 | 0.78 |

**Table 6.6. Classification Consistency by Achievement Level and Cut**

| Grade | Consistency at AL | | | | Consistency at Cut | | | Overall | Kappa |
|---|---|---|---|---|---|---|---|---|---|
| | AL1 | AL2 | AL3 | AL4 | Cut1 | Cut2 | Cut3 | | |
| **Mathematics** | | | | | | | | | |
| 3 | 0.83 | 0.60 | 0.81 | 0.80 | 0.94 | 0.88 | 0.94 | 0.77 | 0.66 |
| 4 | 0.82 | 0.66 | 0.83 | 0.80 | 0.93 | 0.89 | 0.96 | 0.78 | 0.69 |
| 5 | 0.81 | 0.72 | 0.82 | 0.80 | 0.93 | 0.89 | 0.96 | 0.78 | 0.69 |
| 6 | 0.81 | 0.73 | 0.78 | 0.74 | 0.92 | 0.88 | 0.97 | 0.77 | 0.66 |
| 7 | 0.85 | 0.79 | 0.81 | 0.80 | 0.93 | 0.91 | 0.97 | 0.81 | 0.73 |
| 8 | 0.79 | 0.74 | 0.82 | 0.76 | 0.90 | 0.91 | 0.97 | 0.78 | 0.68 |
| HS | 0.65 | 0.43 | 0.64 | 0.80 | 0.80 | 0.81 | 0.96 | 0.60 | 0.45 |
| **Reading** | | | | | | | | | |
| 3 | 0.70 | 0.68 | 0.80 | 0.71 | 0.93 | 0.87 | 0.93 | 0.74 | 0.61 |
| 4 | 0.72 | 0.62 | 0.82 | 0.69 | 0.93 | 0.88 | 0.93 | 0.74 | 0.60 |
| 5 | 0.76 | 0.57 | 0.83 | 0.70 | 0.94 | 0.90 | 0.92 | 0.76 | 0.61 |
| 6 | 0.68 | 0.59 | 0.83 | 0.69 | 0.94 | 0.88 | 0.92 | 0.75 | 0.59 |
| 7 | 0.64 | 0.67 | 0.81 | 0.74 | 0.94 | 0.88 | 0.93 | 0.75 | 0.61 |
| 8 | 0.77 | 0.67 | 0.81 | 0.70 | 0.94 | 0.89 | 0.92 | 0.75 | 0.63 |
| HS | 0.71 | 0.52 | 0.79 | 0.62 | 0.89 | 0.87 | 0.93 | 0.70 | 0.55 |

## 6.2. Fairness and Accessibility

Assessment fairness and accessibility are addressed through multiple approaches in this report. First, Universal Design is used to design the test and items (see Section 2.5.2). 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.

### *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 nonuniform 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 nonuniform if the effect varies conditional on the ability level; that is, nonuniform 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):

Model 1: $logit(P) = \beta_0 + \beta_1 X + \beta_2 E$
Model 2: $logit(P) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 E$
Model 3: $logit(P) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG + \beta_4 E$

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 ($\beta_2$) are related to uniform DIF, and the interaction estimates ($\beta_3$) 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 6.7 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 6.7. LR DIF Categories**

| DIF Category | Level of DIF | Definition |
|---|---|---|
| A | Negligible | $x^2$ test is not significant at 0.05 level or $\Delta R^2$ < 0.035. |
| B | Moderate | $x^2$ test is significant at 0.05 level and 0.035 ≤ $\Delta R^2$ < 0.070. |
| C | Strong | $x^2$ test is significant at 0.05 level and $\Delta R^2$ ≥ 0.070. |

*Note*. $\Delta R^2$ 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 6.8 and Table 6.9 present the numbers 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; a few C DIF items were found per grade in the item pool.

**Table 6.8. DIF Analysis Results—Operational Items**

| Grade | Focal Group | Item Count by DIF Category | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | A | B+ | B- | C+ | C- |
| **Reading** | | | | | | | |
| 3 | Female | 162 | 162 | – | – | – | – |
| | Black or African American | 162 | 145 | 3 | 6 | 1 | 7 |
| | Hispanic/Latino | 162 | 146 | 2 | 7 | 3 | 4 |
| | Asian | 158 | 149 | 5 | 1 | 1 | 2 |
| | Two or More Races | 162 | 158 | – | 1 | 3 | – |
| 4` | Female | 228 | 228 | – | – | – | – |
| | Black or African American | 226 | 217 | 3 | 2 | 2 | 2 |
| | Hispanic/Latino | 228 | 215 | | 4 | 4 | 5 |
| | Asian | 225 | 214 | 3 | 4 | 2 | 2 |
| | Two or More Races | 228 | 226 | 1 | – | – | 1 |
| 5 | Female | 233 | 233 | – | – | – | – |
| | Black or African American | 233 | 229 | 1 | 1 | – | 2 |
| | Hispanic/Latino | 233 | 225 | 1 | 1 | 5 | 1 |

| Grade | Focal Group | Item Count by DIF Category | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | A | B+ | B- | C+ | C- |
| | Asian | 230 | 222 | 3 | 1 | 1 | 3 |
| | Two or More Races | 233 | 230 | 1 | – | 2 | – |
| 6 | Female | 202 | 202 | – | – | – | – |
| | Black or African American | 199 | 186 | 1 | 5 | 3 | 4 |
| | Hispanic/Latino | 202 | 185 | 2 | 5 | 2 | 8 |
| | Asian | 201 | 190 | – | 6 | 3 | 2 |
| | Two or More Races | 202 | 197 | 1 | – | 3 | 1 |
| 7 | Female | 208 | 208 | – | – | – | – |
| | Black or African American | 208 | 200 | 2 | – | 2 | 4 |
| | Hispanic/Latino | 208 | 201 | 1 | 1 | 2 | 3 |
| | Asian | 208 | 202 | 3 | 1 | – | 2 |
| | Two or More Races | 208 | 206 | – | – | 2 | – |
| 8 | Female | 236 | 236 | – | – | – | – |
| | Black or African American | 233 | 213 | 3 | 7 | 3 | 7 |
| | Hispanic/Latino | 234 | 212 | 2 | 8 | 4 | 8 |
| | Asian | 229 | 218 | 1 | 4 | 2 | 4 |
| | Two or More Races | 234 | 228 | 3 | – | 2 | 1 |
| HS | Female | 47 | 47 | – | – | – | – |
| | Black or African American | 47 | 47 | – | – | – | – |
| | Hispanic/Latino | 47 | 47 | – | – | – | – |
| | Asian | 47 | 43 | 3 | – | 1 | – |
| | Two or More Races | 47 | 47 | – | – | – | – |
| **Mathematics** | | | | | | | |
| 3 | Female | 346 | 346 | – | – | – | – |
| | Black or African American | 345 | 340 | 2 | – | 1 | 2 |
| | Hispanic/Latino | 346 | 343 | 1 | – | | 2 |
| | Asian | 346 | 341 | 2 | – | 2 | 1 |
| | Two or More Races | 346 | 345 | – | – | 1 | – |
| 4 | Female | 361 | 361 | – | – | – | – |
| | Black or African American | 361 | 354 | 3 | – | 1 | 3 |
| | Hispanic/Latino | 361 | 356 | 1 | 1 | 2 | 1 |
| | Asian | 359 | 348 | 1 | 3 | 4 | 3 |
| | Two or More Races | 361 | 360 | – | – | – | 1 |
| 5 | Female | 364 | 364 | – | – | – | – |
| | Black or African American | 364 | 359 | 1 | 1 | 2 | 1 |
| | Hispanic/Latino | 364 | 357 | 2 | 1 | 2 | 2 |
| | Asian | 363 | 359 | – | 2 | 1 | 1 |
| | Two or More Races | 364 | 363 | – | – | – | 1 |
| 6 | Female | 324 | 324 | – | – | – | – |
| | Black or African American | 324 | 315 | – | 5 | 3 | 1 |
| | Hispanic/Latino | 324 | 318 | 1 | 2 | – | 3 |
| | Asian | 322 | 316 | 2 | 1 | 1 | 2 |

| Grade | Focal Group | Item Count by DIF Category | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | A | B+ | B- | C+ | C- |
| | Two or More Races | 324 | 323 | – | 1 | – | – |
| 7 | Female | 329 | 329 | – | – | – | – |
| | Black or African American | 328 | 322 | 1 | 1 | 4 | – |
| | Hispanic/Latino | 329 | 324 | 1 | – | 2 | 2 |
| | Asian | 326 | 314 | 3 | 1 | 5 | 3 |
| | Two or More Races | 329 | 328 | – | – | 1 | – |
| 8 | Female | 339 | 339 | – | – | – | – |
| | Black or African American | 336 | 331 | 1 | 2 | 2 | – |
| | Hispanic/Latino | 339 | 334 | 1 | 2 | 2 | – |
| | Asian | 337 | 328 | 2 | 2 | 1 | 4 |
| | Two or More Races | 339 | 337 | 2 | – | – | – |
| HS | Female | 63 | 63 | – | – | – | – |
| | Black or African American | 63 | 59 | – | 3 | 1 | – |
| | Hispanic/Latino | 63 | 50 | 1 | 3 | 8 | 1 |
| | Asian | 62 | 58 | 2 | – | 2 | – |
| | Two or More Races | 63 | 61 | – | – | – | 2 |

**Table 6.9. DIF Analysis Results—Field Test Items**

| Grade | Focal Group | Item Count by DIF Category | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | A | B+ | B- | C+ | C- |
| **Reading** | | | | | | | |
| 3 | Female | 5 | 5 | – | – | – | – |
| | Black or African American | 5 | 5 | – | – | – | – |
| | Hispanic/Latino | 5 | 5 | – | – | – | – |
| | Asian | 5 | 4 | – | – | – | – |
| | Two or More Races | 5 | 5 | – | – | – | – |
| 4 | Female | 5 | 5 | – | – | – | – |
| | Black or African American | 5 | 5 | – | – | – | – |
| | Hispanic/Latino | 5 | 4 | – | – | – | – |
| | Asian | 5 | 5 | – | – | – | – |
| | Two or More Races | 5 | 5 | – | – | – | – |
| 5 | Female | 7 | 7 | – | – | – | – |
| | Black or African American | 7 | 7 | – | – | – | – |
| | Hispanic/Latino | 7 | 7 | – | – | – | – |
| | Asian | 7 | 7 | – | – | – | – |
| | Two or More Races | 7 | 7 | – | – | – | – |
| 6 | Female | 5 | 5 | – | – | – | – |
| | Black or African American | 5 | 5 | – | – | – | – |
| | Hispanic/Latino | 5 | 4 | 1 | – | – | – |
| | Asian | 5 | 5 | – | – | – | – |
| | Two or More Races | 5 | 5 | – | – | – | – |
| 7 | Female | 5 | 5 | – | – | – | – |

| Grade | Focal Group | Item Count by DIF Category | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | A | B+ | B- | C+ | C- |
| | Black or African American | 5 | 5 | – | – | – | – |
| | Hispanic/Latino | 5 | 5 | – | – | – | – |
| | Asian | 5 | 4 | – | – | 1 | – |
| | Two or More Races | 5 | 5 | – | – | – | – |
| 8 | Female | 12 | 12 | – | – | – | – |
| | Black or African American | 12 | 9 | 2 | – | – | 1 |
| | Hispanic/Latino | 12 | 8 | 1 | 1 | | 2 |
| | Asian | 12 | 10 | – | 1 | 1 | – |
| | Two or More Races | 12 | 12 | – | – | – | – |
| HS | Female | 117 | 117 | – | – | – | – |
| | Black or African American | 117 | 108 | 3 | – | 2 | 4 |
| | Hispanic/Latino | 117 | 114 | – | – | 2 | 1 |
| | Asian | 114 | 102 | 5 | 1 | – | 6 |
| | Two or More Races | 117 | 109 | 2 | – | 3 | 3 |
| **Mathematics** | | | | | | | |
| 3 | Female | 6 | 6 | – | – | – | – |
| | Black or African American | 6 | 6 | – | – | – | – |
| | Hispanic/Latino | 6 | 6 | – | – | – | – |
| | Asian | 6 | 6 | – | – | – | – |
| | Two or More Races | 6 | 6 | – | – | – | – |
| 4 | Female | 11 | 11 | – | – | – | – |
| | Black or African American | 11 | 11 | – | – | – | – |
| | Hispanic/Latino | 11 | 11 | – | – | – | – |
| | Asian | 11 | 11 | – | – | – | – |
| | Two or More Races | 11 | 11 | – | – | – | – |
| 5 | Female | 5 | 5 | – | – | – | – |
| | Black or African American | 5 | 4 | – | – | 1 | – |
| | Hispanic/Latino | 5 | 5 | – | – | – | – |
| | Asian | 5 | 4 | 1 | – | – | – |
| | Two or More Races | 5 | 5 | – | – | – | – |
| 6 | Female | 18 | 18 | – | – | – | – |
| | Black or African American | 18 | 18 | – | – | – | – |
| | Hispanic/Latino | 18 | 18 | – | – | – | – |
| | Asian | 18 | 18 | – | – | – | – |
| | Two or More Races | 18 | 18 | – | – | – | – |
| 7 | Female | 19 | 19 | – | – | – | – |
| | Black or African American | 19 | 19 | – | – | – | – |
| | Hispanic/Latino | 19 | 19 | – | – | – | – |
| | Asian | 19 | 19 | – | – | – | – |
| | Two or More Races | 19 | 19 | – | – | – | – |
| 8 | Female | 13 | 13 | – | – | – | – |
| | Black or African American | 13 | 13 | – | – | – | – |

| Grade | Focal Group | Item Count by DIF Category | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total | A | B+ | B- | C+ | C- |
| | Hispanic/Latino | 13 | 13 | – | – | – | – |
| | Asian | 13 | 13 | – | – | – | – |
| | Two or More Races | 13 | 13 | – | – | – | – |
| HS | Female | 145 | 145 | – | – | – | – |
| | Black or African American | 145 | 139 | 2 | – | 3 | 1 |
| | Hispanic/Latino | 145 | 142 | 1 | – | 1 | 1 |
| | Asian | 144 | 135 | 2 | 1 | 2 | 4 |
| | Two or More Races | 145 | 142 | 1 | – | – | 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. Constructing the 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 Commissioner of Education (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 6.10 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 6.10. Available Not-Tested Codes**

| NTC Code | Description |
|---|---|
| INV | **Invalid:** Student's assessment was invalidated, such as due to a security breach. |
| EMW | **Emergency Medical Waiver:** Student was not assessed because of an approved emergency medical waiver. |
| RMV | **Removal:** Student appears in the Acacia platform but is no longer eligible for assessment (for example, due to moving out of state). |

## 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 by the completion of the test. In other words, an adaptive test has multiple test forms by nature.

## 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 2024 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 2024 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. *Maine Through Year Assessment Checklist Spring 2024 Administration*
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 Manage Online Testing Guide*
6. *The Maine Through Year Assessment User and Student Management Guide*
7. *The Maine Through Year Assessment Accessibility Guide*
8. *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 all students (who are eligible to test) participate in the state assessments.

## 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. 13).

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 “[School Administrative Units] 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” (p. 14). 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 2.5.2 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, multi-select, and composite (which uses elements of both multiple choice and multi-select).

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 do not 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 universal tools is 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 numbers 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. Numbers of Students Who Used TTS**

| Grade | Content Area | Number of Students |
|---|---|---|
| 3 | Math | 2,780 |
| 3 | Reading | 2,327 |
| 4 | Math | 2,846 |
| 4 | Reading | 2,411 |
| 5 | Math | 2,562 |
| 5 | Reading | 2,188 |
| 6 | Math | 1,961 |
| 6 | Reading | 1,612 |
| 7 | Math | 1,759 |
| 7 | Reading | 1,596 |
| 8 | Math | 1,472 |
| 8 | Reading | 1,348 |
| HS | Math | 498 |
| HS | Reading | 457 |

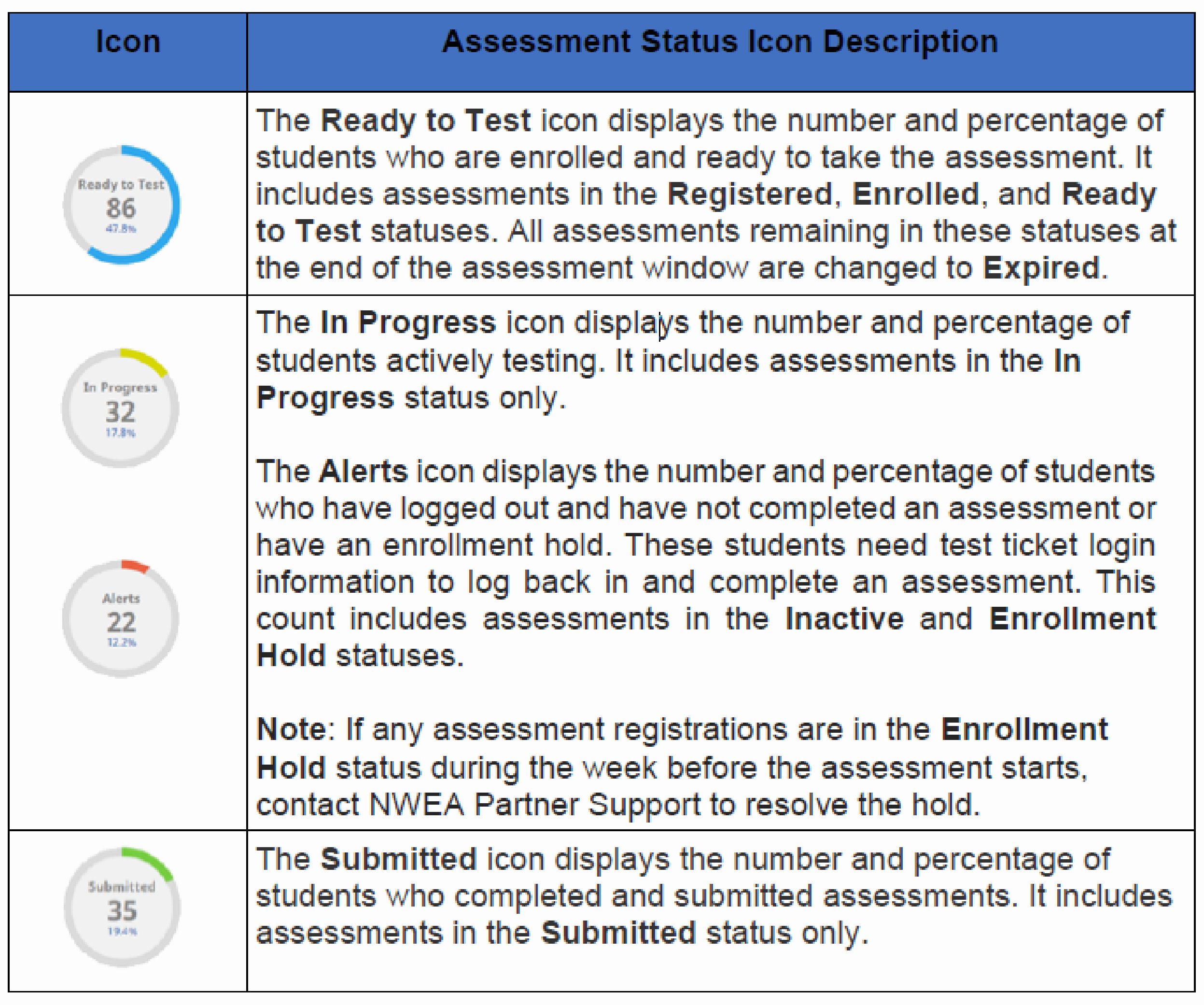
## 7.1. Monitoring Test Administration for Special Populations

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

### *7.1.1. 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 |
|---|---|
| Ready to Test<br>86<br>47.8% | 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**. |
| In Progress<br>32<br>17.8%<br>Alerts<br>22<br>12.2% | The **In Progress** icon displays the number and percentage of students actively testing. It includes assessments in the **In Progress** status only.<br>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.<br>**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. |
| Submitted<br>35<br>19.4% | The **Submitted** icon displays the number and percentage of students who completed and submitted assessments. It includes assessments in the **Submitted** status only. |

### *7.1.2. Maine DOE Site Visits*

In May 2024, 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 2024 *Maine Through Year Assessment Observation Form* shown in Figure 7.2.

**Figure 7.2. 2024 Maine Through Year Assessment Observation Form**

Items indicated in ***bold italic font*** are areas of focus for the Spring 2024 Maine Through Year Assessment.

| | |
|---|---|
| ***School Name:*** | |
| Assessment Administrator: | Proctor/TA/AA(s): |
| ***Observer:*** | ***Subject:*** |
| ***Date of Observation:*** | ***Grade:*** |

| | **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 describe student performance across four 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, the Achievement Level Descriptors (ALDs) shown in Table 2.13 were developed early in the test-development lifecycle to support domain definition (e.g., explication of the construct of interest), item development, and standard setting.

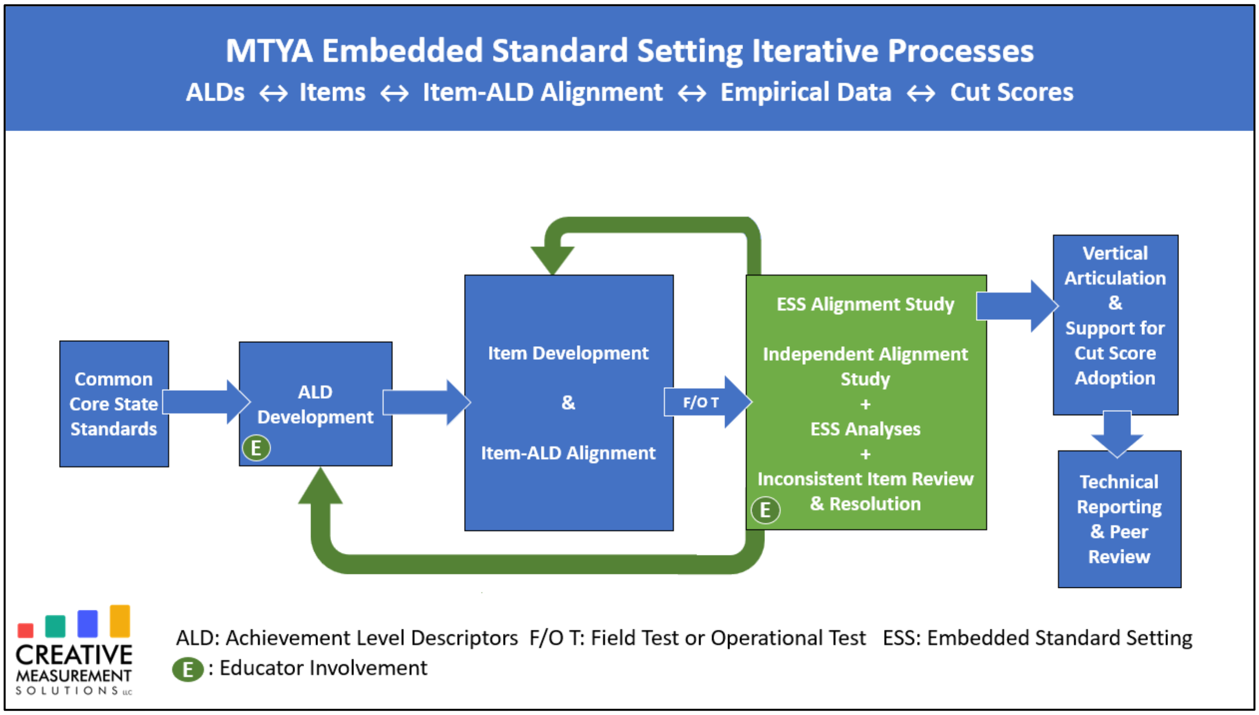
Three cut scores were adopted, defining the four levels of achievement:

- 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 MTYA 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 Achievement Level Descriptors (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. The standard setting technical report is provided in Appendix L.

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


MTYA Embedded Standard Setting Iterative Processes
ALDs ↔ Items ↔ Item-ALD Alignment ↔ Empirical Data ↔ Cut Scores
Common Core State Standards
ALD Development
E
Item Development & Item-ALD Alignment
F/O T
ESS Alignment Study
Independent Alignment Study + ESS Analyses + Inconsistent Item Review & Resolution
E
Vertical Articulation & Support for Cut Score Adoption
Technical Reporting & Peer Review
CREATIVE MEASUREMENT SOLUTIONS LLC
ALD: Achievement Level Descriptors F/O T: Field Test or Operational Test ESS: Embedded Standard Setting
E : Educator Involvement

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 ALDs. ESS processes include:

- **Achievement Level Descriptor development:** This is an articulation of the intended interpretations of the Maine Through Year Assessment across the achievement levels.
- **The ESS Alignment Study:** This is 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 MTYA 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.

In support of this iterative process, NWEA reviewed the results of the July 2023 alignment study and identified a recommended plan of action that was initially presented to the Maine DOE in April 2024. This plan included action items to add additional information into the assessment blueprint to ensure clarity, to undertake a focused review the ALD language based on feedback provided in the alignment study, and to identify/develop additional items for Spring 2025 field testing based on an item bank review and analysis. NWEA will consult with the Maine DOE as this work progresses.

For reading and mathematics, the adopted cut scores were presented to the Commissioner of Education and were approved on August 28, 2023. Table 8.1–Table 8.4 present the final approved cut scores that were used for scoring and the associated impact data.

**Table 8.1. Final Approved Cut Scores—Reading**

| Grade | Cut Scores | | |
|---|---|---|---|
| | ***Below State Expectations*** | ***At State Expectations*** | ***Above State Expectations*** |
| 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 8.2. Impact Data Associated with Cut Scores—Reading**

| Grade | Percent at Level | | | |
|---|---|---|---|---|
| | ***Well-Below State Expectations*** | ***Below State Expectations*** | ***At State Expectations*** | ***Above State Expectations*** |
| 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 8.3. Final Approved Cut Scores—Mathematics**

| Grade | Cut Scores | | |
|---|---|---|---|
| | ***Well-Below State Expectations*** | ***Below State Expectations*** | ***At State Expectations*** |
| 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 8.4. Impact Data Associated with Cut Scores—Mathematics**

| Grade | Percent at Level | | | |
|---|---|---|---|---|
| | ***Well-Below State Expectations*** | ***Below State Expectations*** | ***At State Expectations*** | ***Above State Expectations*** |
| 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% |
| 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*

Achievement Level Descriptors (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 (see Section 2.4 and Table 2.13 for more details about ALDs). 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 MTYA, 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 MTYA, reports were developed and are available at the district, school, group, and individual student levels. Table 8.5 presents a description of each report. A more detailed report explanation can be found in Appendix F.

**Table 8.5. Report Levels**

| Report Name | Aggregation Level | Summary |
|---|---|---|
| District Report | District (SAU) | 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 and 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 |

| Report Name | Aggregation Level | Summary |
|---|---|---|
| Individual Student Report (Spring Only) | Individual Student | Shows all tests in all available content areas for a student in this academic year; designed for parents and families |
| RIT Report | Varies—based on user type | Shows RIT score information for all students matching the search criteria, including RIT score, achievement percentile, and reporting category RIT |
| 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 |
| Comparison Summary Report | School | Shows aggregate comparison of multiple organizations by grade, subject, and student demographics |
| Student Results File | District and State | Downloadable export of student-level data at district and state levels during the test window |

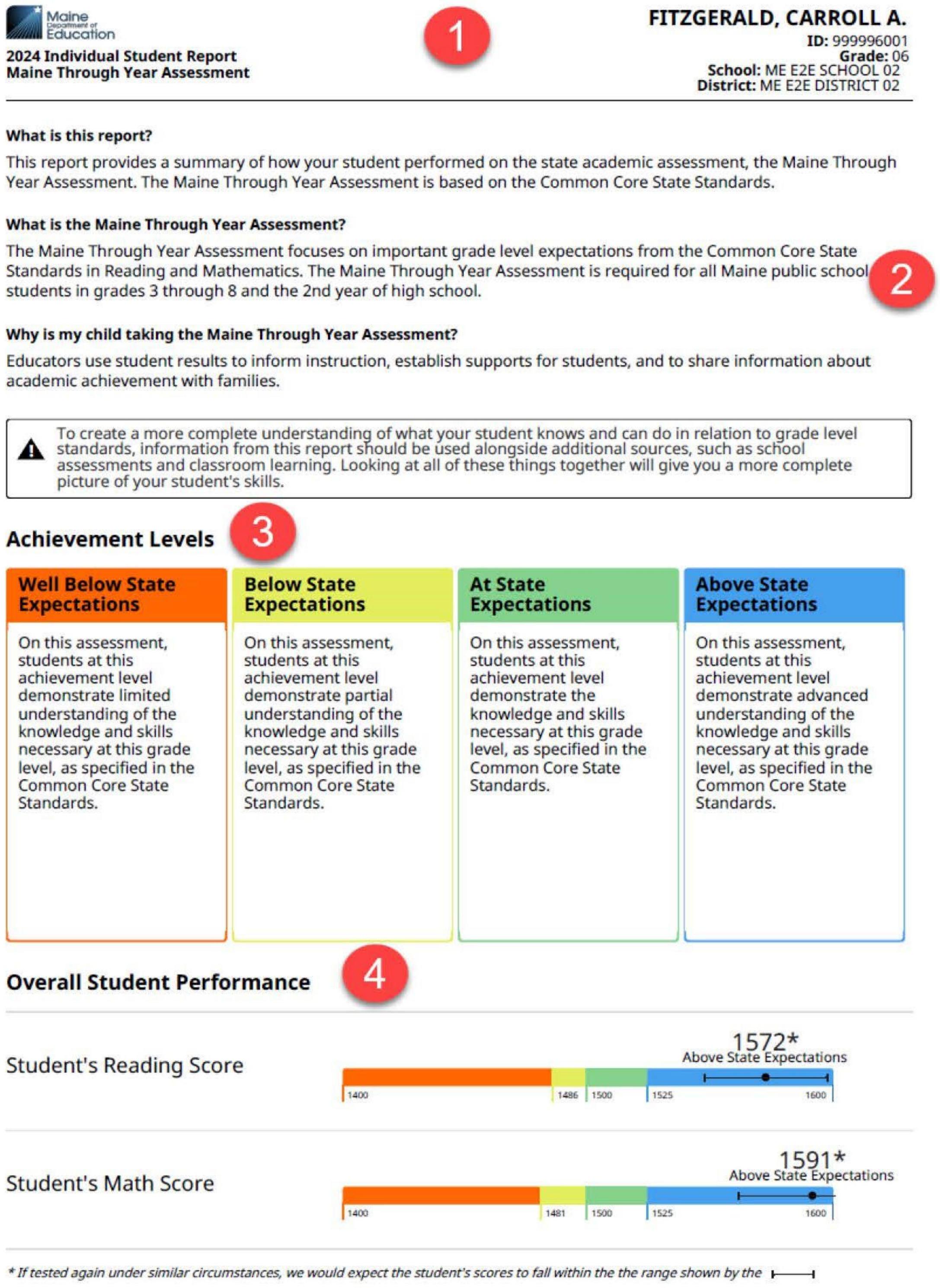
Figure 8.2 shows a mockup of the Individual Student Report (ISR). The ISR is a one-page report designed to show a student's achievement on the Maine Through Year reading and mathematics assessments to parents and families.

In November 2023, a group of Maine educators worked collaboratively with Maine DOE to develop parent-friendly, accessible language and formatting for the Maine Through Year Assessment ISRs. In addition, this educator panel created supplemental documents to explain the reading and mathematics skills focused on at each grade level in easy-to-understand terms. These Individual Student Report supplemental pages for families have been translated into Maine's top ten languages. English examples can be found in Appendix K.

Educators can print ISRs in batches, making them easy to distribute after testing is complete.

The ISRs are generated for the spring assessment and will not be available for the fall and winter assessments.

**Figure 8.2. Individual Student Report**



Maine Department of Education

**2024 Individual Student Report**
**Maine Through Year Assessment**

**FITZGERALD, CARROLL A.**
**ID:** 999996001
**Grade:** 06
**School:** ME E2E SCHOOL 02
**District:** ME E2E DISTRICT 02

**What is this report?**

This report provides a summary of how your student performed on the state academic assessment, the Maine Through Year Assessment. The Maine Through Year Assessment is based on the Common Core State Standards.

**What is the Maine Through Year Assessment?**

The Maine Through Year Assessment focuses on important grade level expectations from the Common Core State Standards in Reading and Mathematics. The Maine Through Year Assessment is required for all Maine public school students in grades 3 through 8 and the 2nd year of high school.

**Why is my child taking the Maine Through Year Assessment?**

Educators use student results to inform instruction, establish supports for students, and to share information about academic achievement with families.

⚠ To create a more complete understanding of what your student knows and can do in relation to grade level standards, information from this report should be used alongside additional sources, such as school assessments and classroom learning. Looking at all of these things together will give you a more complete picture of your student's skills.

**Achievement Levels**

| 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. | 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. | 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. | 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. |

**Overall Student Performance**

Student's Reading Score

1572*
Above State Expectations

Student's Math Score

1591*
Above State Expectations

** If tested again under similar circumstances, we would expect the student's scores to fall within the the range shown by the* ⊢—⊣

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

# References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* AERA. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

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: Issues and Practice*, *39*(1), 8–21. https://doi.org/10.1111/emip.12318

Linacre, J. M. (2015). *Winsteps®* (Version 3.90.2) [Computer software]. 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.]. 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). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 734–755). Praeger.

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). *Probabilistic models for some intelligence and attainment tests*. MESA Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press.

Samejima, 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*. 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