



MAP Growth Course-Specific Tests Technical Report for 2024–2025

May 2026

MAP Growth Course-Specific Tests Technical Report for 2024–2025

© 2026 HMH Education Company. NWEA, MAP, MAP Growth, and MAP Reading Fluency are registered trademarks of HMH Education Company in the U.S. and other countries. All rights reserved.

Technical Report Summary

MAP Growth course-specific tests are computer adaptive assessments administered during the fall, winter, spring, and summer testing terms. The course-specific offerings include Algebra 1, Algebra 2, Geometry, Integrated Math 1, Integrated Math 2, Integrated Math 3, and high school Biology/Life Science. These assessments are designed to measure student achievement in defined course-based content areas and to provide information about students' knowledge and skills relative to the content measured in each course. Scores are reported on the RIT scale, an equal-interval scale that supports interpretation of student performance and comparison across administrations. By adapting item difficulty to student performance during testing, MAP Growth course-specific tests provide efficient and precise measurement while maintaining alignment to targeted course content.

This document is a comprehensive technical report examining MAP Growth course-specific assessments administered during the 2024–2025 school year. It is organized around the primary claims of the assessment. Each chapter is titled after a specific claim, and chapter contents provide evidence to support that claim. The claims addressed in the report are listed below, along with a summary of the evidence provided in each chapter.

Claim: Test Development Is Grounded in Modern Psychometrics

Test development is grounded in modern psychometrics, particularly item response theory (IRT), which provides advantages over classical test theory. IRT enables the ordering of the domain from easier to more-difficult elements and places item difficulty and person ability on the same continuous scale, facilitating adaptive testing and score comparability across different item sets. The Rasch model's property of specific objectivity supports an equal-interval scale for measuring growth. Items are calibrated to the RIT scale, and scoring uses maximum likelihood estimation with fencing to handle extreme response patterns. The item pool is carefully constructed to represent the content domain, aligned to standards where applicable, and organized into instructional areas with proportional representation based on standards emphasis.

Claim: The Target Domain Is Defined, Ordered, and Represented by the Item Pool

A target domain is defined by a set of content standards adopted by an educational agency. Examples include state content standards, Common Core State Standards, Next Generation Science Standards, and the NWEA Content Frameworks. Instructional areas are established according to content standards and reflect the articulation of content within a specific high school-level course. The assessments also include content that serves as prerequisites to the content within the course. Once the test design is established, items are selected from a large item bank to create a test-specific item pool that contains only items aligned to the content standards of interest. Visualizations in this chapter illustrate the way standards are mapped to instructional areas. Content specialists review items for quality, clarity, alignment, cognitive complexity, and appropriateness.

Claim: Item Development Follows a Rigorous Process

MAP Growth assessments incorporate a range of item interaction types (selection, construction, generation, and composite items) to measure the breadth of content standards and span the range of cognitive complexity. Item development begins with a standards-driven item acquisition plan informed by item pool analyses and areas of need. Detailed item specifications guide

writers to ensure alignment to targeted standards, appropriate cognitive complexity, and adherence to best practices in item construction. Throughout development, items are supported by extensive metadata (e.g., grade, DOK, Bloom's taxonomy, provisional RIT, allowable tools) that facilitate pool management, test assembly, and internal analyses. Items undergo multiple layers of review—including content accuracy, bias and sensitivity, accessibility, copyright and permissions, and adherence to formatting and style guidelines.

Following content and editorial review, new items are embedded within operational test events for field testing. Psychometric calibration uses a fixed-person item design, with item difficulty estimated on the RIT scale using a proportional curve-fitting approach. Items must meet explicit statistical criteria—including acceptable p values, discrimination indices, Rasch fit statistics, and distractor functioning—to become operational. Operational items are continuously monitored for parameter drift and content quality.

Claim: Test Administration Is Standardized and Secure

MAP Growth test administration is standardized and secure, with proctors controlling test sessions and accommodations assigned based on student needs. The testing platform supports millions of student test events annually, with certifications for high concurrent user loads. Security measures include encrypted data transmissions, secure data centers with redundancy, lockdown browsers to prevent unauthorized access, and procedures to prevent test content exposure. Large item pools and adaptive testing reduce the risk of content sharing among students. Access to the system is role-based, ensuring appropriate permissions for test setup, administration, data management, and privacy.

Claim: Test Events Sample the Domain and Adapt in a Principled Way

MAP Growth employs a computer adaptive testing algorithm that selects items based on a student's estimated ability during testing. In general, correct responses lead to more difficult items, while incorrect responses lead to easier items. At the same time, content targeting and prioritization ensures that each test meets test blueprint requirements by selecting items from across the instructional areas. The Enhanced Item Selection Algorithm prioritizes course-aligned content that targets the typical grade-level of the course. Simulation studies and analyses of operational data indicate that the adaptive engine produces tests that satisfy blueprint requirements while supporting precise measurement of student achievement.

Claim: Student Scores and Item Difficulty Are on a Stable Vertical Scale

Score reliability is evaluated through internal consistency and score stability over time. For MAP Growth course-specific assessments, internal consistency is estimated using marginal reliability within the IRT framework, while test-retest reliability reflects the extent to which scores remain stable across testing intervals. Marginal reliability coefficients are consistently high, ranging from 0.94 to 0.98 across courses and administrations, and test-retest reliability shows the expected pattern of stronger relationships over shorter intervals (winter–spring) than longer intervals (fall–spring). Measurement error is quantified by the conditional standard error of measurement (CSEM), which remains generally stable across ability levels, with slightly higher values at the lower and upper extremes.

Claim: Scores Align with Their Intended Purpose, Use, and Interpretation

MAP Growth scores provide norm-referenced and criterion-referenced information to support instructional decision-making. User norms allow comparison among a large sample of test takers, while linking studies allow prediction of proficiency on state summative tests. The Learning Continuum organizes skills and concepts represented in the item pool by difficulty along the RIT scale. This can help inform decisions about where to begin formative assessment, highlight areas where students may benefit from enrichment or additional scaffolding, and guide instructional priorities.

Claim: The Test Is Fair for All Examinees

MAP Growth is designed to be fair for all students, incorporating universal design principles, designated features, and accommodations to reduce construct-irrelevant variance and ensure equitable access to the measured construct. Item content is reviewed for bias and sensitivity, with differential item functioning (DIF) analyses indicating that the vast majority of items show negligible DIF across gender and racial/ethnic groups. MAP Growth includes embedded and non-embedded accessibility features, including text-to-speech, screen readers, braille devices, bilingual dictionaries, and various assistive technologies. Test engagement is monitored to detect rapid guessing, with proctor alerts to re-engage students and maintain score validity.

Claim: Score Reports Facilitate the Interpretation and Use of Scores

A variety of reports are available at the student, class, school, and district levels, allowing test score users to find information appropriate for their level of decision-making (e.g., student, school, or district level). For example, the Student Profile report offers individualized learner data, growth tracking, instructional area scores, and proficiency projections, whereas the School Profile report includes aggregates of student scores by each grade in a school. Professional learning resources support educators in interpreting and applying MAP Growth data effectively.

In conclusion, the evidence presented in this report supports the use of MAP Growth course-specific assessments as psychometrically sound, fair, accessible, and instructionally useful measures of student achievement across diverse populations and educational contexts.

Table of Contents

1. Introduction	1
1.1. A Brief History of MAP Growth.....	1
1.2. MAP Growth Overview and Theory of Action	1
1.2.1. Comprehensive Assessment System	2
1.2.2. MAP Growth Assessment Design	5
1.2.3. Target Population and Test Fairness.....	5
1.2.4. Intended Uses of Results	6
2. Test Development Is Grounded in Modern Psychometrics	9
2.1. Traditional Assessment	9
2.2. Item Response Theory	11
3. The Target Domain Is Defined, Ordered, and Represented by the Item Pool	15
3.1. The Item Pool Is Aligned to Content Standards	16
3.2. The Item Pool Is Ordered by Grade Level.....	23
3.3. Test Blueprints Represent the Content Standards	24
4. Item Development Follows a Rigorous Process	27
4.1. Item Specifications	27
4.1.1. Metadata	28
4.2. Item Content Alignment	28
4.2.1. Alignment to Standards	28
4.2.2. Cognitive Complexity	29
4.3. Item Content Review	29
4.3.1. Copyright and Legal Permissions Review	30
4.3.2. Content Validation: Accuracy and Relevance Review	30
4.3.3. Item Owner: In-Depth Item Review.....	30
4.3.4. Item Quality Review	32
4.4. Field Testing, Calibration, and Psychometric Review	32
4.5. Item Types.....	33
5. Test Administration Is Standardized and Secure	39
5.1. Test Engagement Functionality	39
5.2. Administration Training	40
5.3. Practice Tests.....	40
5.4. Test Security	40
5.4.1. Data Security	41
5.4.2. Role-Based Access.....	41
6. Test Events Sample the Domain and Adapt in a Principled Way.....	43

6.1. Computer Adaptive Test Administration.....	43
6.1.1. Statistical Requirements.....	43
6.1.2. Content Requirements	43
6.1.3. Enhanced Item Selection Algorithm.....	45
6.1.4. MAP Growth Item Selection	46
6.2. Simulation Procedures and Test Publishing.....	48
6.3. Test Blueprints Have Been Satisfied.....	49
6.4. The Domain Is Sampled for Efficient Tests and Aggregate Statistics	52
7. Student Scores and Item Difficulty Are on a Stable Vertical Scale.....	53
7.1. Operational Item Statistics	53
7.1.1. Classical Item Difficulty and Item Discrimination.....	53
7.2. Item Fit	57
7.2.1. Cross-Sectional Item Fit.....	57
7.2.2. Longitudinal Item Fit.....	57
7.3. Examinee Fit	59
7.4. Examinee Scores and Item Difficulty Increase by Grade and Show Variability.....	61
7.5. Conditional Standard Error of Measurement.....	61
7.5.1. CSEM Results.....	62
7.6. Score Reliability.....	66
7.6.1. Marginal Reliability Results	66
7.6.2. Test-Retest Reliability Results.....	69
8. Scores Align with Their Intended Purpose, Use, and Interpretation.....	71
8.1. Scale Scores	71
8.1.1. Momentary RIT Scores	71
8.1.2. Final RIT Scores	71
8.1.3. Instructional Area Scores	72
8.2. MAP Growth Norms.....	72
8.3. Linking Studies and Predicted Proficiency	72
8.3.1. Linking Study Methodology	73
8.3.2. Linking Study Accuracy in Predicting Proficiency	73
8.4. Relationships with Measures of the Same Construct	74
8.5. The Learning Continuum	75
8.5.1. Overview of the Learning Continuum and Its Development	75
8.5.2. Relationship of the Learning Continuum to the RIT Scale and Student Scores	75
8.5.3. Instructional Uses of the Learning Continuum	75
9. The Test Is Fair for All Examinees	77

9.1. Test Taker Demographics	77
9.2. Universal Design	80
9.3. Accommodations	80
9.3.1. Universal Features	81
9.3.2. Designated Features	82
9.3.3. Accommodations.....	83
9.3.4. Third-Party Assistive Software	84
9.4. Differential Item Functioning	84
9.4.1. Mantel-Haenszel Procedure and ETS DIF Classification Levels.....	85
9.4.2. MAP Growth DIF Results	86
10. Score Reports Facilitate the Interpretation and Use of Scores	90
10.1. Student Profile Report	90
10.2. Family Report	93
10.3. Class Reports	93
10.3.1. Class Profile Report	94
10.3.2. Achievement Status and Growth Report.....	95
10.4. School and District Level Reports	96
10.4.1. School Profile Report	96
10.4.2. District Profile Report	101
10.5. Score Interpretation Guide.....	103
10.6. Professional Learning	103
10.7. On-Demand Learning	105
10.8. Additional Resources for Developing Understanding	105
10.8.1. Online Help Center.....	105
10.8.2. NWEA Connection	105
10.8.3. Resource Center	106
10.8.4. NWEA YouTube Channel.....	106
References	107
Appendix A: Marginal Reliability by Course, State, and Term.....	110
Appendix B: Marginal Reliability by Course, Instructional Area, and Term	112

List of Tables

Table 3.1. Item Difficulty Summary Statistics for All Items in the CCSS/NGSS Item Pools	24
Table 3.2. Example Test Blueprints.....	25
Table 4.1. Examples of Item Metadata Types	28
Table 4.2. Item Review Checklist	31
Table 4.3. Field Test Item Review Criteria.....	33
Table 5.1. Test Security Before and During Testing	41
Table 5.2. Access Roles, Permissions, and Responsibilities	42
Table 6.1. Simulation Study Evaluation Points	49
Table 6.2. Percentage Distribution of Blueprint Alignment Effect Sizes by Course and Term	51
Table 7.1. Summary of Item <i>P</i> Values.....	55
Table 7.2. Summary of Item Point-Biserial Correlations	56
Table 7.3. Item Misfit Summary by Course.....	59
Table 7.4. Comparison of Examinee RIT Scores to Item Difficulties	61
Table 7.5. CSEM Summary by Course, Term, and Score Decile.....	64
Table 7.6. CSEM Summary by Course, Demographic Group, Term, and Score Decile	64
Table 7.7. Marginal Reliability by Course and Term	67
Table 7.8. Marginal Reliability by Course, Demographic Group, and Term	67
Table 7.9. Test-Retest Reliability by Course	69
Table 7.10. Test-Retest Reliability by Course and Demographic Group	69
Table 8.1. Classification Accuracy for MAP Growth Course-Specific Tests and State Summative Assessments	74
Table 8.2. Correlations Between MAP Growth Course-Specific Tests and State Summative Assessments	75
Table 9.1. Gender Distribution by Course and Term	77
Table 9.2. Race/Ethnicity Distribution by Course and Term.....	79
Table 9.3. Universal Design for Learning Principles	80
Table 9.4. Available Universal Features	81
Table 9.5. Available Designated Features.....	82
Table 9.6. Available Accommodations	83
Table 9.7. Third-Party Assistive Software.....	84
Table 9.8. DIF Classification Levels	86
Table 9.9. DIF Classification Summary by Course, Term, and Group	87
Table 10.1. Professional Learning Workshops	104

List of Figures

Figure 1.1. Theory of Action Infographic.....	4
Figure 3.1. Alluvial Diagram of an Item Pool for Algebra 1	17
Figure 3.2. Alluvial Diagram of an Item Pool for Algebra 2	18
Figure 3.3. Alluvial Diagram of an Item Pool for Geometry	19
Figure 3.4. Alluvial Diagram of an Item Pool for Integrated Math 1	20
Figure 3.5. Alluvial Diagram of an Item Pool for Integrated Math 2	21
Figure 3.6. Alluvial Diagram of an Item Pool for Integrated Math 3	22
Figure 3.7. Alluvial Diagram of an Item Pool for Biology/Life Science.....	23
Figure 4.1. Item Development Flowchart.....	27
Figure 4.2. Multiple-Choice (Science)	34
Figure 4.3. Multiple-Choice (Math)	34
Figure 4.4. Multiselect/Multiple Select (Math).....	35
Figure 4.5. Hot Text/Selectable Text (Math).....	35
Figure 4.6. Drag-and-Drop/Click-and-Click (Science).....	36
Figure 4.7. Click-and-Pop (Math)	36
Figure 4.8. Text Entry (Math)	37
Figure 4.9. Composite Item (Math).....	37
Figure 4.10. Composite Item (Science)	38
Figure 7.1. Item OUTFIT Statistics by Course and Term	58
Figure 7.2. Examinee OUTFIT Statistics by Course and Term	60
Figure 7.3. CSEM Distribution by Course and Score Decile	63
Figure 8.1. Learning Continuum Example	76
Figure 10.1. Example of Student Profile Report	91
Figure 10.2. Example of Subject Scores and Highlights Sections of Student Profile Report	91
Figure 10.3. Example of Comparisons, Instructional Areas, and Growth Goals Sections of Student Profile Report	92
Figure 10.4. Example of Growth Over Time Section of Student Profile Report	92
Figure 10.5. Example of Family Report	93
Figure 10.6. Example of Class Profile Report.....	94
Figure 10.7. Example of Achievement Status and Growth Report.....	95
Figure 10.8. Grouping Students Based on Performance in Achievement Status and Growth Report	96
Figure 10.9. Example School Profile Report.....	97
Figure 10.10. Example of Single-Term Achievement Tab in School Profile Report	98
Figure 10.11. Example of Growth and Achievement Tab in School Profile Report	99
Figure 10.12. Example of Growth and Achievement Quadrant by Grade in School Profile Report	100
Figure 10.13. Example of Growth and Achievement by Grade in School Profile Report.....	101
Figure 10.14. Example of Single-Term Achievement Tab in District Profile Report	102
Figure 10.15. Example of Growth and Achievement Tab in District Profile Report	103
Figure 10.16. Professional Learning Online	105

1. Introduction

1.1. A Brief History of MAP Growth

NWEA was founded in 1976 by a consortium of school districts seeking practical, efficient, and accurate ways to measure student achievement and growth. Founders included Allan Olson, George S. Ingebo, and Vic Doherty. With the later contributions of Susan Smoyer, Ron Houser, and Gage Kingsbury, NWEA transitioned from paper-and-pencil testing to the first computer adaptive test in 1985.

Originally named the Northwest Evaluation Association, the organization later became known simply as NWEA. In 2023, NWEA was acquired by HMH, formerly Houghton Mifflin Harcourt. Since its founding, NWEA has focused on developing assessments that provide educators with accurate, timely information about student learning while minimizing testing time. NWEA is best known for MAP Growth, but its assessment portfolio also includes MAP Reading Fluency, course-specific assessments, and state assessment programs that integrate interim and summative solutions.

Measures of Academic Progress (MAP) was officially introduced in 2000 as the first large-scale, standardized computerized adaptive testing program for K–12 students in the United States. Initially, MAP served students in grades 3–12 across subjects including Reading, Language Usage, Mathematics, General Science, and Concepts and Processes; the Concepts and Processes assessment was later retired. MAP assessments were designed to use adaptive item selection and vertically scaled scores to support comparisons of student achievement across terms and grade levels. In 2006, NWEA expanded its offerings with MAP for Primary Grades to provide developmentally appropriate reading and math assessments for students in kindergarten through grade 2. NWEA began offering course-specific tests for math in 2017 and for science in 2018. Over time, Measures of Academic Progress was rebranded as MAP Growth, and MAP for Primary Grades became MAP Growth K–2.

Early MAP implementations required testing software to be installed locally on school computers and administered over local school networks. By the late 2000s, as school internet infrastructure improved, MAP Growth transitioned to a fully web-based platform, supporting more streamlined administration, broader accessibility, and more efficient delivery of assessment updates. In 2025, NWEA fully implemented another notable revision to MAP Growth through the Enhanced Item Selection Algorithm (EISA).

While MAP Growth is designed to measure achievement and growth across broad subject domains using vertically scaled assessments, MAP Growth course-specific assessments are designed to measure achievement in specific secondary-level courses. These assessments extend NWEA's adaptive assessment approach to course-based contexts, providing educators with information about student performance in subjects such as Algebra, Geometry, Integrated Math, and Biology/Life Science. The course-specific assessments are intended to support interpretation of student achievement within defined course content while maintaining the psychometric quality, efficiency, and instructional usefulness associated with the MAP Growth assessment system.

1.2. MAP Growth Overview and Theory of Action

MAP Growth is an interim assessment that may be administered in fall, winter, spring, and summer. Test scores are reported on the Rasch Unit (RIT) scale, a vertical scale that supports interpretation of student achievement and growth across administrations. MAP Growth course-specific tests extend this approach to high school math and science courses by measuring

achievement in defined course-based content areas such as Algebra 1, Algebra 2, Geometry, Integrated Math 1, Integrated Math 2, Integrated Math 3, and Biology/Life Science.

MAP Growth course-specific assessments are computer adaptive tests that account for both statistical and content requirements. The adaptive engine selects items with difficulty values that match a student's estimated ability level while prioritizing course-aligned content needed to satisfy the test blueprint. Through this design, course-specific tests provide information about student achievement at a point in time and, when administered across terms, changes in achievement over the course of the year. These results help districts, schools, teachers, and families understand student performance relative to the content measured in specific high school courses.

The MAP Growth theory of action describes the role of MAP Growth within a comprehensive assessment system. Its central premise is that when MAP Growth is used for its intended purposes—to provide information about student learning and support data-informed decisions about curriculum and instruction—it can contribute to improved learning for all students. Figure 1.1 illustrates the components of the MAP Growth theory of action and the relationships among the claims that support the validity argument. Evidence for these claims is presented throughout this technical report and in related research reports.

1.2.1. Comprehensive Assessment System

A comprehensive assessment system involves multiple forms of assessment that each emphasizes a different purpose and use (Perie et al., 2009). Summative, interim, and formative assessments complement one another and provide the appropriate data for decision-making at every level of an education agency. MAP Growth is one such component of a comprehensive assessment system.

A key outcome in education is for students to demonstrate proficiency with the curriculum established for their grade. Summative assessments are a measure of this outcome. They involve a sample of tasks from a broadly defined curriculum and a determination of a student's level of proficiency. Summative assessments are tightly aligned to standards, and test administration follows highly standardized procedures. They are typically given at the end of the school year, and results are usually available after instruction for the year has completed. Consequently, summative assessments have very limited utility for measuring student progress during the year or informing instructional decisions. MAP Growth is *not* a summative assessment, but it may be used to predict performance on a summative assessment, allowing educators to identify students at risk of not meeting state proficiency standards by the end of the year.

Although MAP Growth is not a summative assessment in and of itself, it has been tightly integrated with summative assessments in several states. In these innovative programs, MAP Growth tests are given in fall and winter. A short summative assessment is then given in the spring, supplemented with additional MAP Growth items so that a RIT score is produced along with a summative scale score. In this manner, MAP Growth scores are provided for all three terms, and an additional summative score is provided in the spring. This enables the measure of growth and proficiency from the same assessment program.¹

¹ Summative testing programs that incorporate MAP Growth RIT scores are described in detail in separate technical reports.

MAP Growth is a vertically scaled interim assessment that provides information about student achievement during a particular academic term as well as changes in student achievement (i.e., growth) over multiple terms. The comparability of scores across multiple administrations is a particular strength of MAP Growth. Comparability is achieved through a test design that incorporates grade-level standards and cross-grade adaptability to meet students where they are in their learning and precisely measure their achievement and growth. MAP Growth is administered up to four times per year (fall, winter, spring, and summer). Both the frequency of the assessment and the comparability of scores are useful for measuring growth, setting goals for students, planning instruction for a term, and evaluating programs; however, the assessment is less useful at providing more frequent feedback or daily tailoring of instruction to student needs.

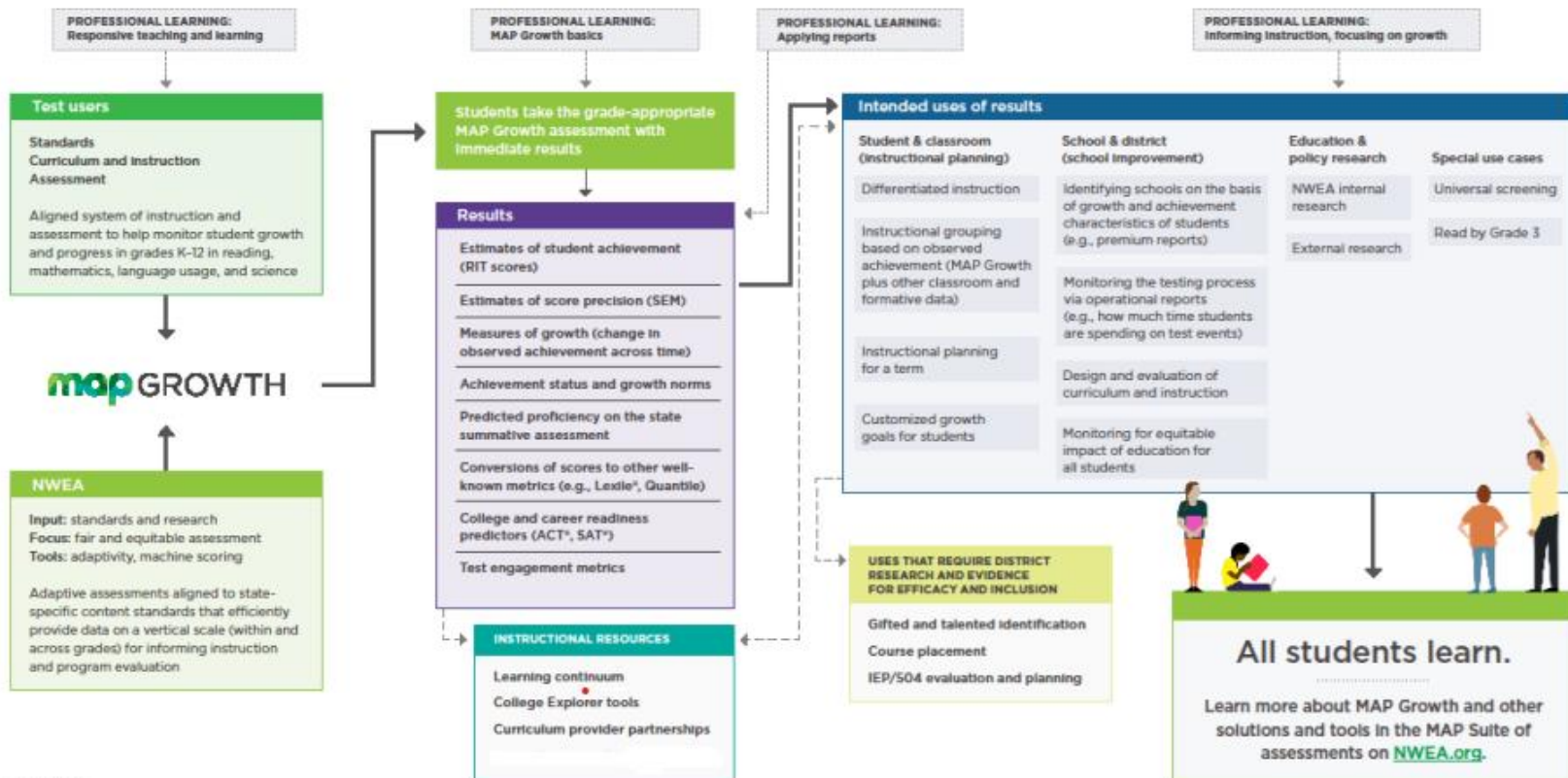
Alternatively, formative assessments are frequent instructional practices aligned to specific lesson plans or instructional units that facilitate the tailoring of instruction on a regular basis, perhaps even daily, and adapting instruction to student needs. MAP Growth is not a formative assessment, but NWEA supports formative instructional practices through professional learning in responsive teaching and learning (see Nordengren, 2023, for more detail). In the professional learning programs, teachers learn to create high-quality formative assessments to support their day-to-day instruction. They also learn to combine that information with MAP Growth scores to make data-informed decisions. Formative assessments are also possible through HMH digital curriculum offerings, such as the Math 180 and Read 180 programs and the Waggle product. MAP Growth scores are leveraged for initial placement into these products, and the regular formative assessments within them monitor student progress.

Figure 1.1. Theory of Action Infographic

MAP Growth theory of action

MAP® Growth™ measures achievement and growth in K-12 math, reading, language usage, and science. The research-backed assessment provides teachers with accurate and actionable evidence to help target instruction for students working on, above, or below grade level.

The MAP Growth theory of action illustrates the how and why of MAP Growth—from test development to leveraging results. With every test, for every use, our goal is clear: work alongside educators to help all students learn.



1.2.2. MAP Growth Assessment Design

MAP Growth course-specific tests are aligned to content standards to support measurement within defined high-school course contexts. Instructional areas are established according to standards and reflect the articulation of content within a specific high school-level course. The assessments also include content that serves as prerequisites to the content within the course. Once the test design is established, items are selected from a large item bank to create a test-specific item pool that contains only items aligned to the content standards of interest. Each item pool includes multiple item types such as basic multiple-choice items and more sophisticated technology-enhanced items. Item pools are deep enough to implement longitudinal exposure constraints and make certain that a student never sees the same item more than once during a long period of time. Every test must undergo a computer adaptive test simulation of four consecutive test administrations to ensure that the item pool and item development plans will support the intended test design and satisfy performance metrics.

MAP Growth course-specific scores are reported on the Rasch Unit (**Rasch unit**) scale that ranges from 100 to 350 with a mean of 200 and a standard deviation of 10. The RIT scale has two key characteristics. First, it is a vertical scale with equal-interval units. For course-specific tests, the vertical scale supports score comparisons across administrations within the same content area (math or science) and, where appropriate, across related course sequences within each content area. For example, Algebra 1 scores from a fall administration can be interpreted on the same math scale as Algebra 2 scores from a spring administration. Second, the equal-interval property of the RIT scale allows differences in RIT scores to be interpreted consistently across the scale, supporting the measurement of student growth over time.

MAP Growth course-specific tests are designed to measure student achievement efficiently and precisely in high school math and science courses. Because the assessments are adaptive, item difficulty is tailored to student performance during testing, which supports measurement precision across a wide range of ability levels. A fixed-form assessment of similar length would generally provide less targeted measurement, while achieving comparable precision would typically require more items and additional testing time. In addition, each student's score is reported with an associated estimate of measurement error.

MAP Growth course-specific assessments can be administered in-school or remotely. In both settings, test sessions are managed by a trained proctor who can assist students with technical issues and other testing needs. During in-school administrations, the proctor is physically present with students; during remote administrations, proctoring is facilitated through video conferencing, chat, email, or other communication channels. Non-adaptive practice tests are available online to help students become familiar with MAP Growth question formats and item types.

Test security procedures address both the integrity of the assessment and the privacy of student data. Security measures include role-based access, data encryption, system redundancy across geographically distributed locations, and other safeguards. During administration, a lockdown browser may be used to limit access to unauthorized resources, and proctor monitoring helps ensure that students complete the assessment independently.

1.2.3. Target Population and Test Fairness

Three aspects of fairness in testing are (a) fairness in treatment during the testing process, (b) fairness as a lack of measurement bias, and (c) fairness in access to the measured construct (AERA et al., 2014). MAP Growth course-specific assessments address these aspects of

fairness through item and test design, accessibility features and accommodations, bias and sensitivity review, statistical analyses of item functioning, and procedures intended to reduce threats to valid score interpretation.

MAP Growth course-specific assessments are designed to provide fair measurement for students taking high school math and science courses. The adaptive nature of the assessments helps match item difficulty to each student's estimated ability level. As a result, students across the achievement continuum receive items that are appropriately challenging, which supports efficient measurement and helps produce comparable levels of measurement precision across students.

Fairness is also addressed through item-development and review procedures. Items are reviewed to ensure that content is appropriate, inclusive, and free from language or contexts that could unfairly advantage or disadvantage students because of race, ethnicity, gender, cultural background, or other student characteristics unrelated to the measured construct. These qualitative reviews are complemented by statistical analyses designed to detect potential item bias. Items flagged through statistical procedures are reviewed further and may be revised or removed from the item pool.

Universal design features and accommodations are incorporated to support access for a wide range of students. Embedded accessibility tools may include features such as volume amplification, a line reader, zoom or screen magnification, and other supports. Non-embedded and designated supports may include tools or administration conditions such as a Spanish dictionary, text-to-speech, color contrast, human reader, native language translation, or a separate testing setting. Additional accommodations, including assistive technology, extended time, screen reader, and refreshable braille, may be provided for students with an Individualized Education Program, 504 plan, or other documented accommodation need.

Student engagement is another consideration for fair and valid measurement. Students who are more engaged in testing tend to obtain scores that more accurately reflect their achievement than students who are disengaged (Wise & DeMars, 2005). Proctors are trained to administer test sessions in ways that support student engagement, and MAP Growth includes functionality to detect possible disengagement during testing. When low engagement is detected, the system can prompt the student and alert the proctor so that appropriate intervention can occur. These procedures are intended to help students give their best effort and obtain scores that more accurately reflect their knowledge and skills in the assessed course content.

1.2.4. Intended Uses of Results

The intended uses of results in a comprehensive assessment system may be classified as instructional, predictive, or evaluative (Perie et al., 2009). The particular use depends, in part, on the user and the level at which decisions are made. MAP Growth reports provide information to support decision-making across multiple levels of an educational agency, from the individual students to classrooms, schools, districts, and states. Individual student scores may be aggregated to summarize performance for larger groups, but the level of aggregation should align with the intended decision. For example, classroom-level decisions should be based on student-level information and classroom summaries, school-level decisions should be based on classroom and school summaries, and district-level decisions should be based on school- and district-level summaries.

A variety of MAP Growth reports are available, each designed to support different interpretive and instructional purposes. Reports provide student-level scores as well as aggregate summaries at multiple levels of decision-making. The MAP Growth [reports portfolio](#) summarizes available reports and provides guidance for their intended use. In addition, premium reports provide a fine-grained level of detail, such as information about student performance on grade-level standards.

Instructional uses of MAP Growth course-specific scores most often occur in classrooms and involve teachers, students, and families. For individual students, scores provide information about what they likely know and can do within a specific course-based content area at a particular point in time. When students test across multiple terms, results can also provide information about changes in achievement over the course of the year. Students may use this information to reflect on their learning and set goals, while teachers may use it to monitor progress, identify areas of strength and need, and plan instruction. When interpreted alongside classroom evidence, formative assessment results, and other local data, MAP Growth course-specific scores can support differentiated instruction and targeted learning activities.

NWEA provides [instructional connections](#) to supplemental and intervention content providers to support educators in personalizing learning for students based on their MAP Growth scores. These connections include integration with [HMH Performance Suite and HMH Personalized Path](#) as well as over 30 other instructional content providers, including popular providers such as Khan Academy (see <https://blog.khanacademy.org/learning-paths-informed-by-map-growth/>), IXL, Edmentum, and Newsela. These resources typically use MAP Growth scores to place students in learning content and curricula offered by other companies, eliminating the need for additional in-platform testing. The wide availability of these instructional connections is another way that MAP Growth encourages fairness in access to the measured construct.

Many instructional uses of MAP Growth are supported by NWEA professional learning opportunities, as shown in Table 10.1. Modules cover formative assessment practices that are independent of MAP Growth and are based on principles of assessment literacy. Other modules involve MAP Growth-specific training on a variety of topics, including the interpretation of score reports and setting growth goals.

MAP Growth reports also provide norm- and criterion-referenced information to give context to scores and facilitate score interpretation. MAP Growth norms for course-specific tests provide a way to compare student performance with a representative sample of test takers for the fall, winter, and spring terms (NWEA, 2022). Achievement status norms provide percentiles for scores at a particular point in time for each grade level. It is a snapshot of student performance. Growth norms describe a student's change in test scores relative to a national population. In particular, the Conditional Growth Percentile (CGP) describes a student's performance compared with similar students with comparable prior test scores.

MAP Growth course-specific norms use a multilevel growth model that has test administration nested within individual students. Norms allow for a student's achievement status and within-year (i.e., within-course) growth to be compared with similar students.

Criterion-referenced information comes in the form of predictions for summative assessments and connections to external measures. NWEA [linking studies](#) connect MAP Growth course-specific test scores to state summative assessments (Hu, 2021). These studies support predictions of whether a student is on track to meet proficiency expectations on a state end-of-course assessment. For course-specific assessments, linking evidence can help educators

interpret student performance in relation to corresponding high school state assessments, such as Algebra 1, Geometry, or Biology end-of-course assessments. This information can support instructional planning, early intervention, and monitoring of readiness before state summative testing.

Finally, school administrators, district personnel, state policy makers, and researchers may be interested in evaluating the efficacy of a curriculum, policy, or intervention. MAP Growth scores may serve as a common outcome for these evaluations. For example, MAP Growth data have been used to research summer learning loss and achievement disparities (Atteberry & McEachin, 2021; McNeish & Dumas, 2021), test score gaps between advantaged and disadvantaged students (von Hippel & Hamrock, 2019), the potential disadvantages schools may impose on Asian American students (Yoon & Merry, 2017), and personalized learning (Pane et al., 2015).

2. Test Development Is Grounded in Modern Psychometrics

Measurement theory started with classical test theory and has evolved into modern psychometrics with the use of item response theory (IRT). While aspects of classical test theory remain, modern psychometrics provides a number of advantages for test design, adaptive administration, scaling, psychometric analysis, and score interpretation. This section describes the theoretical distinctions between classical and IRT-based methods and explains why an IRT framework is well suited for MAP Growth course-specific assessments. The technical information presented here provides the foundation for the assessment claims examined in this report and for the empirical evidence used to support those claims.

2.1. Traditional Assessment

Classical test theory and traditional assessment were founded upon the idea of a domain sampling model. A domain consists of all possible items that are admissible for a test form, whether real or imagined. For example, a domain could be all possible items that satisfy a set of content standards. In this paradigm, a test form is constructed to be a mini version of the entire domain. It is a random or representative sample of items from the target domain. There are an infinite number of other test forms that could be constructed from the domain; the actual test form is just one sample from this domain. Test scoring is aimed at making an inference about a student's performance in the domain.

A percent correct score is an observed score calculated as the number of items answered correctly on a test form divided by the total number of items on the form multiplied by 100. Specifically, for $i = 1, \dots, n$ students responding to a set of $j = 1, \dots, J$ items, the percent correct score for dichotomous items is calculated as shown in Equation 1.

$$PC_i = 100 \times \sum_{j=1}^J \frac{x_{ij}}{J} \quad (1)$$

where x_{ij} is an examinee's score (i.e., either 0 or 1) on item j . It is an estimate of the true (percent correct) score²—the score a student would earn by answering every item in the domain. Because a percent correct score is obtained from just one of many possible test forms that could be sampled from the domain, it is just one of many possible estimates of the true score. It is necessary to know how close this percent correct score is to the true score and how much the score would change if the form had been constructed from a different sample of items. To calculate the precision of the percent correct score as an estimate of the true score, a conceptual framework is drawn upon conceptual framework of hypothetically repeating the process of sampling a test from the domain, administering it to an examinee, and calculating a score.

To conceptualize the repeated sampling, everything about the test administration is standardized and fixed except for the sample of items on a test form. Thus, the only aspect of test administration that could change upon repeating the test is the sample of items on the test. Any change in score from one sample to the next is due to the sample of items alone. It is this idea of (hypothetically) repeatedly sampling from the domain and giving a new test to an examinee that gives rise to the concepts of the standard error of measurement (SEM) and reliability. If the process of sampling from the domain were repeated, the test would have a new

² The true percent correct score is also called a "domain score," but that usage is avoided herein to prevent confusion with other meanings of the term.

set of items that also represents the domain. Repeating this process an infinite number of times would yield a distribution of percent correct scores for a student. That distribution is centered (i.e., has a mean value) at the student's true percent correct score. The standard deviation of that distribution is known as the person-specific standard error of measurement (SEM_i). It indicates how close the domain scores are to the true score. The variance of true scores for the population of students is called the true score variance, σ_T^2 . The expected value (average) of the person-specific error variances is the error variance for the population of examinees, $\sigma_E^2 = E[SEM_i^2]$, and the SEM for the population of students is the square root of this value, $SEM = \sqrt{\sigma_E^2}$. Reliability, ρ_{XT}^2 , is then defined as the ratio of true score variance, σ_T^2 , to the observed score variance, $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$, for the populations of examinees, $\rho_{XT}^2 = \sigma_T^2 / (\sigma_T^2 + \sigma_E^2)$. Although the repeated sampling is hypothetical, it is necessary to derive equations for calculating the SEM and reliability from one or more test administrations.

Content validity is closely connected to the idea of domain sampling. As previously discussed, each sample of items from the domain is a test form, and there are many possible test forms that could be created. Content validity is evidence that shows the degree to which a test form represents the entire domain and nothing else. It involves judgement about the definition of the domain (e.g., content standards), the extent to which the blueprints represent the domain, and the alignment of items and test content to the domain. If the test content does not fully represent the domain, the test suffers from construct underrepresentation—there are parts of the domain that are not measured by the test. On the other hand, construct-irrelevant variance occurs when a test measures characteristics that are not part of the domain. Thus, inferences about the domain depend on the degree to which a test covers all aspects of the domain and avoids construct underrepresentation and construct-irrelevant variance.

A classical approach to assessment presents a number of limitations. One limitation is the interpretation of a score. A percent correct score represents *how much* of the domain a student has achieved; however, there is no direction to it. It does not indicate which behaviors have been achieved or where the student is going. As an analogy,³ imagine athletes running hurdles around a track. Unlike an event seen in the Olympics, the hurdles in this race vary in size from really short to very tall and are placed randomly around the track. Each runner's score for this race is calculated as the percentage of hurdles that were cleared (were not knocked over) after a lap. While this score does tell something about the runner, the information it provides is limited. It reveals *how many* hurdles were cleared but not the *height* of the hurdles that were cleared. Perhaps the runner cleared all of the low hurdles. Perhaps it was a mix of low and medium-high hurdles. The score itself does not explain which parts of the domain were mastered (which hurdles were cleared); it only tells how much of it was mastered. The use of these scores for improvement is arguably even more limited. To prepare the runner for the next race under these conditions, training is limited to simply attempting more hurdles in the domain, regardless of their size. The score itself does not provide information about the *type* of hurdle that should be practiced for.

Another limitation of the classical approach is that the percent correct score depends on the difficulty of the items. Even though a person's ability does not change, the selection of items can change the observed score. Scores will increase if easier items are selected, and they will decrease if more-difficult items are selected. That is, person ability and item difficulty are

³ This analogy focuses on the statistical requirements of test design. In particular, it focuses on the relationship between item difficulty and test scores. To extend the analogy to content requirements imagine that the hurdles are made of different materials, such as aluminum, wood, or brick. The composition and construction of the hurdles reflect the content requirements.

completely confounded in the classical model. In the hypothetical sampling framework, each random sample of items results in randomly parallel tests, which means that sampled test forms are equally difficult. In practice, test forms must be equated to ensure that the difficulty of the test forms does not affect student scores.

Scoring aside, equally difficult test forms would also lead to a different experience for test takers. Some would perceive the test to be of average difficulty, low-performing students would view the test as difficult, and high-performing students would view the test as easy. However, an adaptive test based on IRT leads to a more uniform experience across examinees of every ability level.

2.2. Item Response Theory

Item response theory (IRT) is a model-based measurement that overcomes many of the limitations of traditional assessment and classical test theory. With IRT, test design is aimed at defining the *ordered* nature of the domain. It expands on classical ideas by not only defining the domain but also the way elements of the domain are ordered from easy to achieve to more difficult to achieve. While content standards can serve as an ordered description of the domain, the ordering is often limited to the progression of standards from one grade to the next. Within each grade, the domain may be conceptually defined with a learning progression or empirically defined by calibrating items to a vertical scale and using the item locations to describe the scale (e.g., item mapping).

IRT is a family of statistical methods for modeling the response to a test item. The Rasch model (Rasch, 1980) is the simplest and perhaps most robust of these models. It has only two parameters. First, the model has a person parameter, θ , that represents a student's ability (i.e., the IRT-based score). Small values indicate low ability levels and large values represent high ability levels. Second, the model has an item difficulty parameter, δ , where low values indicate easy items and large values indicate difficult items. The model is given by Equation 2.

$$P_{ij}(\theta) = \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)} \quad (2)$$

Person ability and item difficulty parameters are on the same continuous scale. The model describes the interaction between a person and an item. It indicates the probability of a correct response, given that a person with a particular ability level responds to an item with a certain difficulty value. A person with an ability value equal to the difficulty value (i.e., $\theta = \delta$) has a 50% chance of answering the item correctly. If ability is larger than difficulty (i.e., $\theta > \delta$), the probability of answering correctly is higher than 50%. Conversely, if ability is lower than difficulty (i.e., $\theta < \delta$), the probability of answering correctly is smaller than 50%. An examinee's IRT-calculated score, θ , represents *where* the examinee is located along the continuum of achievement. Returning to the analogy from the previous section, it is the location of the runner on the track.

A unidimensional IRT scale represents the ordered nature of the domain as a continuum of achievement. Item calibration is a procedure for estimating the location of *items* on the continuum, whereas test scoring is aimed at estimating the location of *examinees* on the underlying continuum. A scale score is a transformation of the person ability parameter, θ , to a metric that is more user friendly, which includes avoiding negative scores. The transformation is $SS(\theta_i) = \theta_i \times A + B$, where A is a slope (scale) coefficient and B is an intercept (location) coefficient. These coefficients are computed to transform the ability parameter from the logit

metric into a scale score. For MAP Growth, $A = 10$ and $B = 200$ so that RIT scores have a mean of 200 and a standard deviation of 10. The same scale score transformation is applied to the item difficulty parameters to keep person ability and item difficulty on the same scale, $SS(\delta_j) = \delta_j \times A + B$. Some scales also truncate scores using the lowest observable scale score (LOSS) and the highest observable scale score (HOSS). Scale scores below the LOSS are converted to the LOSS, and values above the HOSS are converted to the HOSS. Valid MAP Growth scale scores are limited to a range of 100 to 350. The purpose of a scale score is to establish a common metric for examinees and produce scores that have comparable meaning.

If an IRT approach to test design is applied to the previous track analogy, the key difference to the race is that the hurdles are ordered from lowest to highest around the track rather than being randomly placed around the track. Runners would continue around the track until they could no longer clear a hurdle. Because the hurdles are ordered, the event informs how many hurdles a runner can clear *and* the height of the hurdles a runner can clear. A runner who can clear a 1-foot hurdle but not a 2-foot hurdle has no need to continue the race and attempt 3- and 4-foot hurdles. In a similar way, there is no need to have a runner jump over a 6-inch hurdle when the runner can easily clear a 1-foot hurdle. The ordered nature implies that the runner can easily clear shorter hurdles and is unlikely to clear those higher hurdles. To prepare for the next race, the runner's training should focus on those hurdles that are immediately higher than the last cleared hurdle (1.5–2-foot hurdles). There is no need to train for 1- and 1.5-foot hurdles the runner can easily clear, and there is no need to train for 4-foot hurdles when the runner is not able to clear a 3-foot hurdle. Because of the ordered nature of the hurdles, once it is known where on the track a runner can no longer clear a hurdle, it is then known what the runner is able to do and what the runner should practice next. There is no need for the runner to attempt every hurdle.

Expanding on the analogy, IRT item difficulty parameters are the heights of hurdles on the track. Some are easy (e.g., low hurdles) and some are difficult (e.g., high hurdles). A person's IRT-based score represents how far along the track the runner can go before being unlikely to clear the next hurdle. A significant feature of this analogy is that the runner and the hurdles are measured on the same scale. In IRT, item difficulty and person ability parameters are on the same scale. This is a key difference between IRT and classical measurement. As a result, groups of items can be used at distinct locations to describe different parts of the scale as well as the ordering of the scale (e.g., item mapping). The description can then be applied to students who are located at the same part of the scale as a group of items. The MAP Growth Learning Continuum is an example of this approach to score interpretation (see *Scores Align with Their Intended Purpose, Use, and Interpretation*).

Item response models have properties that make them more useful than those of classical test theory. Parameter invariance is an important property of IRT that applies to the item and person parameters.⁴ Item parameter invariance means that item difficulty values do not depend on the group of students. Any group of students will produce the same item difficulty value.⁵ Likewise, person parameter invariance means that the person ability (IRT-based score) does not depend on the set of items on a test.⁵ Parameter invariance is perhaps *the* reason why computer adaptive testing is possible. Two different students can take two entirely different tests with

⁴ Parameter invariance is a property of the parameters. It is not guaranteed to hold for the parameter estimates. Whether the property holds with the estimates must be tested. For item parameters, these statistical tests are called tests for item parameter invariance and include many of the same methods as tests for differential item functioning.

⁵ This is true up to a linear transformation and within estimation error.

items calibrated to the same scale and their scores will be comparable and on the same scale. Parameter invariance is also a feature that distinguishes IRT from classical test theory. In traditional assessment based on classical test theory, person scores depend on the difficulty of the items. A student will earn a higher score when given a test composed of easy items and a lower score when given a test composed of hard items. However, with parameter invariance, a student will receive the same score when given easy items, difficult items, or any set of items when those items are calibrated to a common scale.

Specific objectivity is a concept that is similar to parameter invariance that has a direct implication for the scale itself. It is a property of the Rasch family of models that distinguishes it from other IRT models (Rasch, 1977; Wright & Stone, 1979). Specific objectivity means that the distance between any two examinees will be the same regardless of item difficulty. Using the logit of Equation 2 applied to two different people, specific objectivity is $(\theta_1 - \delta) = (\theta_2 - \delta) = \theta_1 - \theta_2$. The same is true for the item difficulty parameters—the difference between any two item difficulty parameters is the same regardless of person ability, $(\theta - \delta_1) = (\theta - \delta_2) = \delta_1 - \delta_2$. Specific objectivity is the basis for claims that the Rasch model results in an equal-interval scale and that growth can be measured by computing the difference in IRT-based scores obtained on two different occasions (and with two different tests calibrated to the same scale).

Because these properties hold for the parameters and not for the sample estimates, it is imperative to ensure that the model assumptions are supported, the model fits the data, and the parameters remain stable over time.

Two assumptions of the Rasch model are unidimensionality and local independence. Unidimensionality means that there is only one dimension that influences the probability of a correct answer (i.e., the Rasch model has a single person parameter because of the assumption of unidimensionality). When unidimensionality does not hold, other dimensions affect the probability of a response, and the data will be influenced by factors outside of the measurement model (e.g., construct irrelevant variance).

A related assumption is local independence. This assumption means that once person ability is taken into account, there is no relationship among items. From a practical perspective, local independence means that one item does not influence the response to another item. It is an assumption that can be violated in a number of ways, such as item sets and test speededness.

Equation 2 is the basis for other functions used to describe the quality of items and examinees. It is used for fit statistics that summarize the similarity of observed and expected responses to an item. Two such fit statistics are infit and outfit. Values close to 1 indicate the model fits the data well. Larger values (e.g., above 1.5) indicate poorer model-data fit, and smaller values (e.g., below 0.5) indicate overfitting. Infit and outfit statistics may be computed for items to indicate the degree to which an item fits the data as well as for examinees to indicate the extent to which an examinee's score fits the observed data.

Content validity for a test design based on modern psychometrics and adaptive testing is more complex than that for a traditional assessment. Test blueprints and the item pool must be ordered and aligned to content standards. Individual test events should also align to blueprints within a predetermined degree of tolerance. However, score interpretation is not limited to the content of individual test events. Bock et al. (1997) noted that once a score is estimated and the person's location on the scale is known, items representative of the ordered domain may be used to facilitate score interpretation even if the items were never seen by the student. The

ordered scale itself and reporting scores derived from it are also sources of validity evidence related to test content.

In summary, IRT has several implications for test design, administration, and score reporting. First, the test design should reflect the ordered nature of the domain. Second, all items can be calibrated to a common scale that spans multiple grade levels. Third, the test can be adaptive in terms of difficulty and content—the goal of scoring is to precisely locate a student on the scale and measure growth over time. Fourth, scores are a transformation of the person ability parameter, and different scores can provide different types of information about a student's ability.

3. The Target Domain Is Defined, Ordered, and Represented by the Item Pool

A well-structured set of content standards defines the target domain *and* the ordering of elements within the target domain. Standards define what a student should know and be able to do by the end of each grade level or course. Broadly speaking, elements of the target domain increase in difficulty and cognitive complexity as grade level increases. For example, high school Life Science work is more challenging than eighth-grade work, and so on. At a more specific level, curriculum design establishes the within-grade scope and sequence of individual lessons, activities, and assessments to ultimately lead students to achieve the desired end-of-year outcomes (as in “backward design”). The broad learning outcomes established in content standards apply to every student in a state, but curriculum design can differ from one school district to the next. Each district must consider its student population and the instruction needed to help its students learn and achieve end-of-year outcomes. Thus, the target domain is defined and ordered through a combination of content standards and curriculum design.

Good curriculum design accounts for students prepared for learning at their current grade as well as students who are not prepared for on-grade learning along with those who have moved beyond their grade level. While content standards define end-of-year learning outcomes, not every student will experience on-grade instruction throughout the school year. Low-performing students will experience more instruction in below-grade content than typical students, with the goal being to facilitate rapid learning (i.e., large growth) of targeted prerequisite skills and move the student to on-grade or on-course instruction. This is often done through below-grade instruction and providing additional instructional resources as well as scaffolds for accessing on-grade content. At the other end of the continuum are high-performing students who quickly master on-grade content. They may exhibit low growth because it does not take much to move such a student from their existing performance level to mastery of on-grade content. Indeed, high-performing students may start a grade already at or above on-grade proficiency. Consequently, high-achieving students will likely experience enriched instruction and advanced coursework that may exceed on-grade content (e.g., AP courses). The main point is that while every student in a grade will be held to the same end-of-year learning outcomes, not every student will experience the same curriculum and instruction throughout the year. Thus, the target domain involves the progression of grade-to-grade learning outcomes and the type of curriculum experienced through instruction.

Content standards, curriculum design, instruction, and resources that support instruction (e.g., supplemental instruction) are connected to an assessment through test blueprints. Traditional course-specific summative assessment uses blueprints focused on end-of-course learning. Some summative assessments also focus on course-to-course evolution and student growth from one spring to the next spring. A limitation of a traditional course-specific summative assessment is that it provides no information about the progression of student learning during the school year and focuses strictly on course-level content regardless of the curriculum and instruction experienced by a student.

MAP Growth is aligned to content standards and frameworks. The content is flexible and able to adapt to the curriculum and instruction likely experienced by the student. MAP Growth course-specific blueprints represent the content domain defined by standards in math and science. Course-specific blueprints are specific to the high school-level courses referenced in the test names. The courses are commonly Algebra 1, Algebra 2, Geometry, and Life Science or Biology. This use of course-specific blueprints helps ensure that item content is suitable to the targeted students in terms of complexity and developmental appropriateness. Each blueprint indicates the number of items per instructional area and the maximum number of items on the

test. Test blueprints define the cross-grade instructional areas that represent the top level of standards for the course. Grade-specific standards are mapped to each instructional area, with only those prior-grade standards representing prerequisite skills mapped alongside the standards for the targeted course.

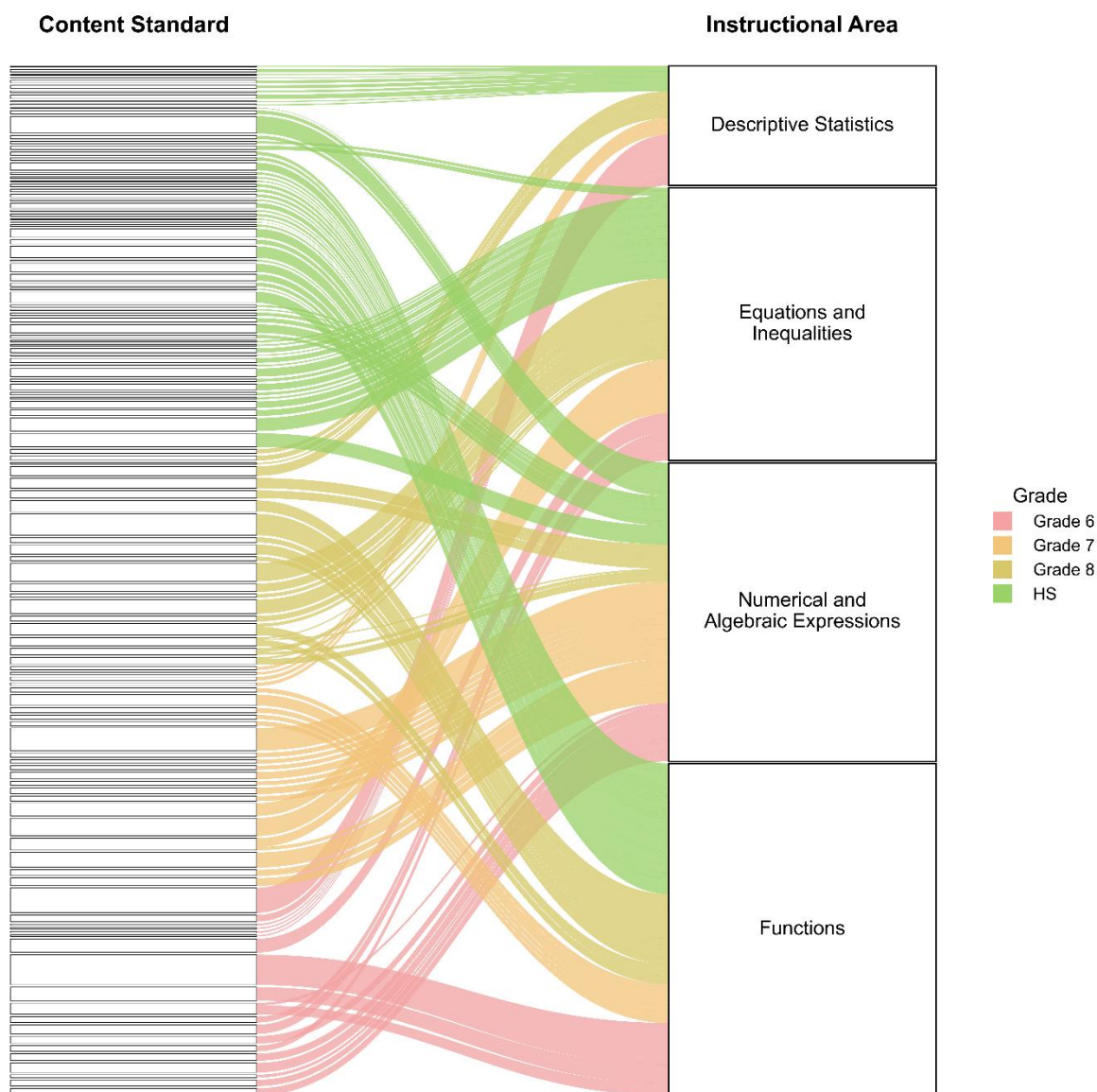
After establishing a set of blueprints, an item pool is created by choosing items from a large item bank and assigning each selected item to a content standard and instructional area. The grade level of an item is assigned according to the grade level of the standard or, if the standard grade is not provided by the standards authority, the standard is assigned the grade(s) at which that content is typically taught. An item is assigned to a range of grades in cases where the standard applies to multiple grades. For example, an item aligned to a grade 6–8 science standard will have an item grade of grade 6, grade 7, and grade 8. The content standards represented in the blueprint go beyond the designated course of the test to ensure that each pool contains items for students who may be performing below the level of the targeted course. Blueprint documents provide specific details for each MAP Growth assessment. Although NWEA conducts its own alignment studies when creating an item pool, its work is often verified through external alignment studies.

3.1. The Item Pool Is Aligned to Content Standards

A MAP Growth item pool is developed for each course-specific test. Items are selected from an existing bank and aligned to content standards to create the item pool for each test.

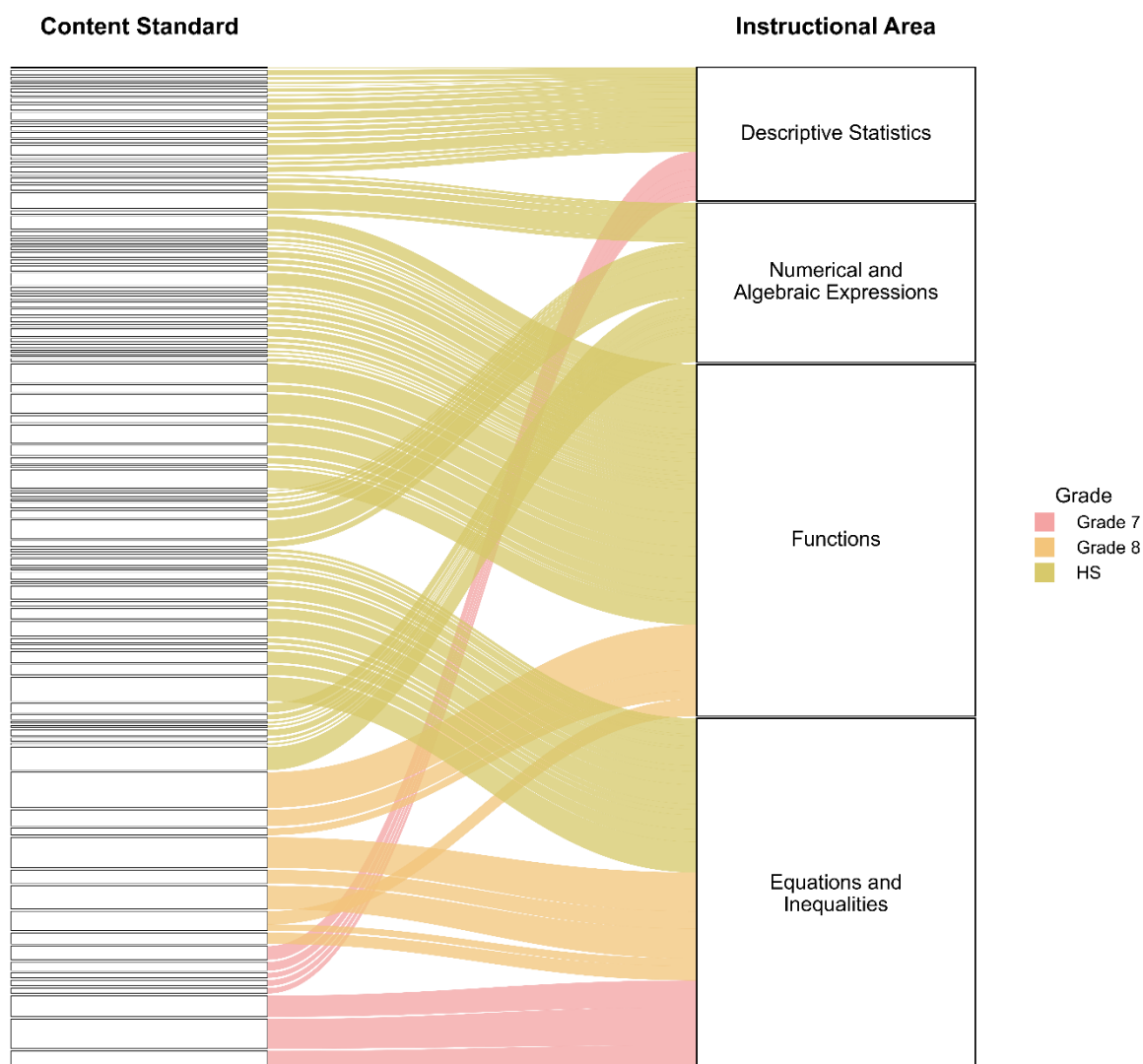
Figure 3.1 through Figure 3.7 illustrate how each course-specific test's content standards map to instructional areas, with the size of each element shown in proportion to the number of items associated with each standard and instructional area. The size of the white boxes on either side of the diagrams is in proportion to the number of items. The larger the white box, the more items for that content standard or instructional area. The content standard names are omitted from the left side of the diagrams to reduce visual clutter. The alluvia flowing between the content standards and instructional areas are colored according to the grade level of the standard, and the thickness of each alluvium represents the number of items.

Figure 3.1. Alluvial Diagram of an Item Pool for Algebra 1



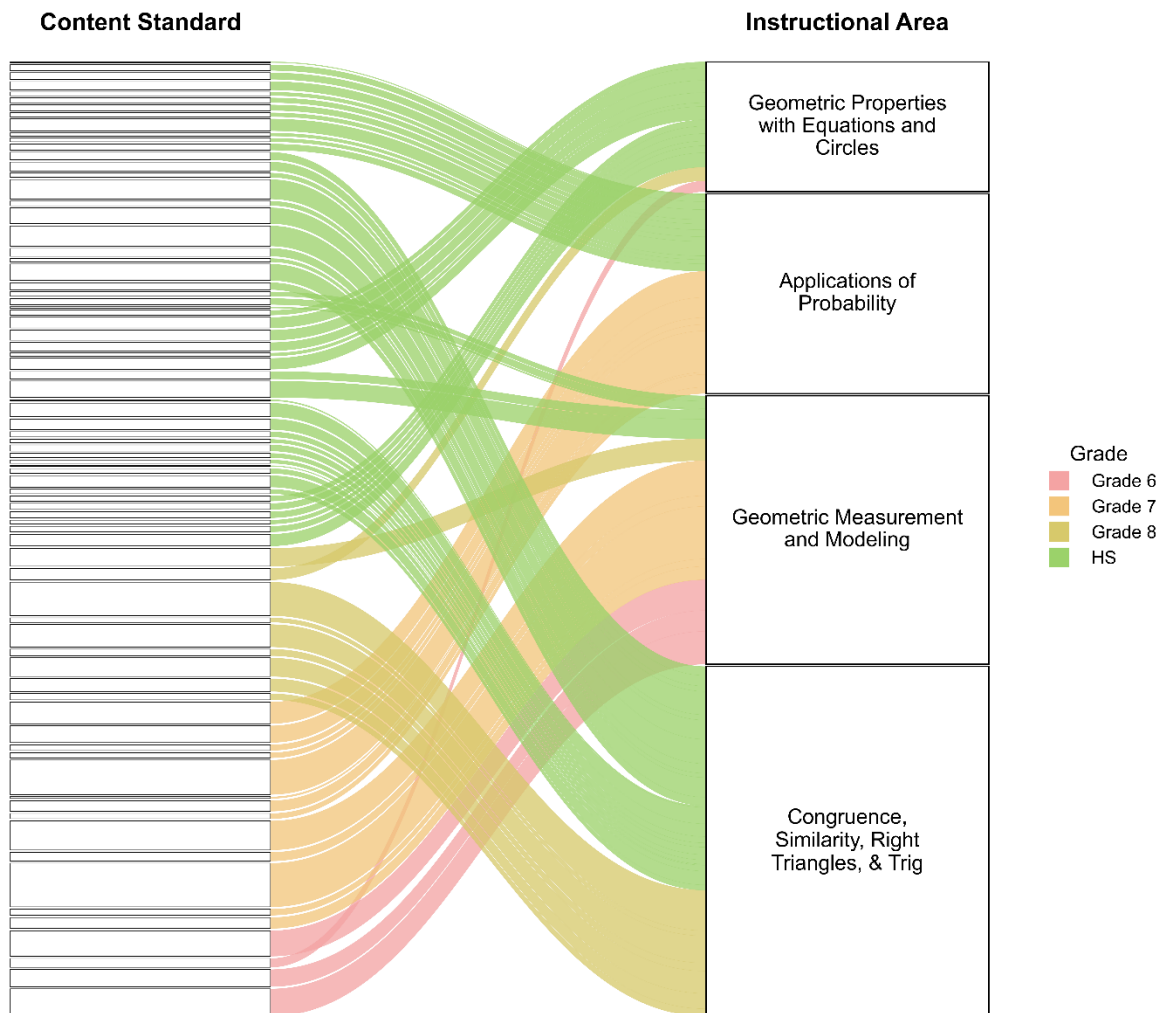
As shown in Figure 3.1, Functions has the greatest number of items in the Algebra 1 item pool, while Descriptive Statistics has the fewest. This design choice is based on the greater prevalence or greater emphasis of those elements in the content standards. The figure also illustrates that prerequisite grade levels are represented in each instructional area. The representation is proportional to the amount of off-grade content devoted to each instructional area. For example, grade 7 has more items in Numerical and Algebraic Expressions than grades 6 and 8 do.

Figure 3.2. Alluvial Diagram of an Item Pool for Algebra 2



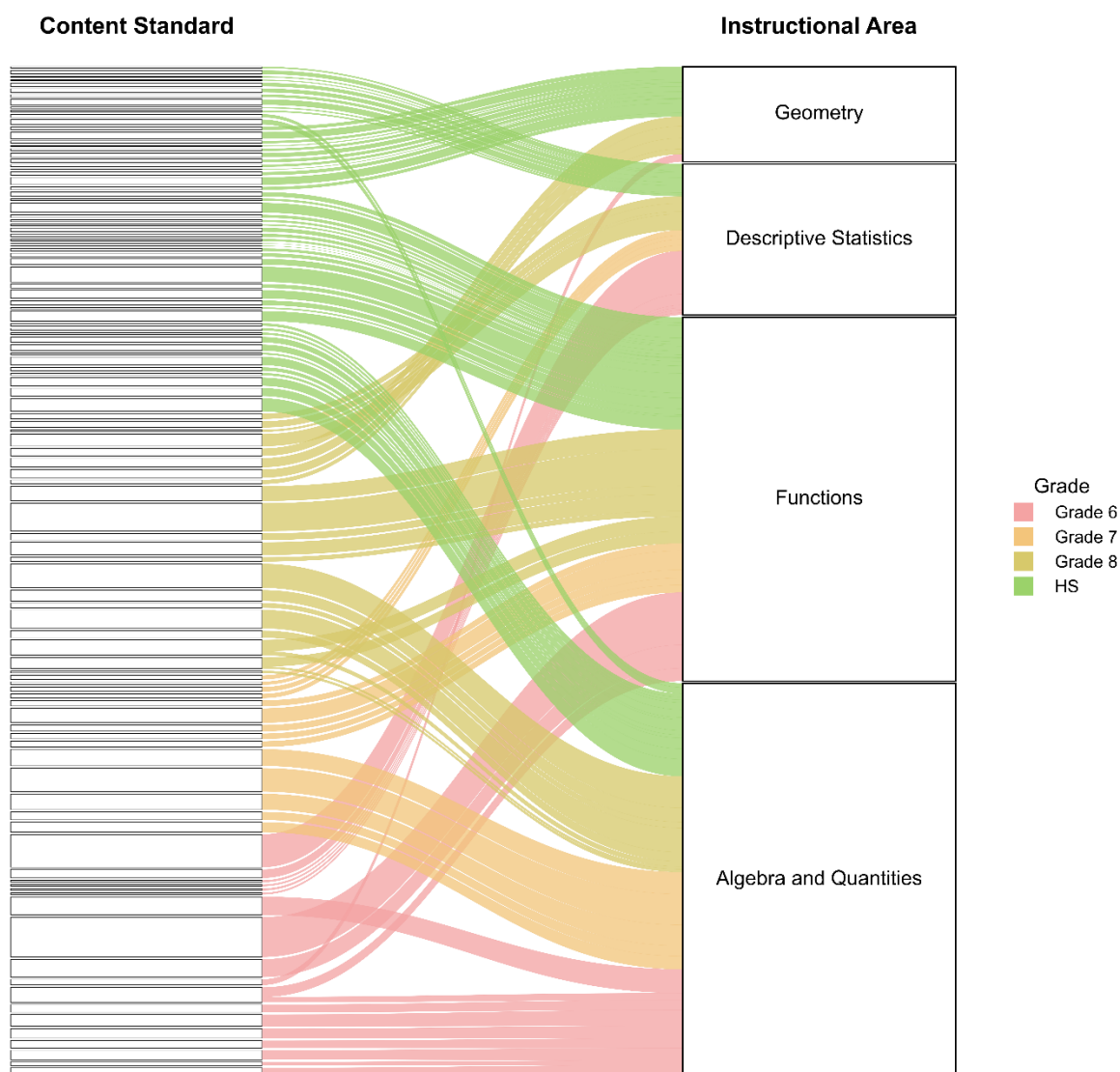
The pool for Algebra 2 shown in Figure 3.2 has the same instructional areas as the Algebra 1 item pool (Figure 3.1). Prerequisite grade levels are represented in instructional areas as appropriate. The Equations and Inequalities and Functions instructional areas contain the greatest number of items, as they represent the larger number of standards.

Figure 3.3. Alluvial Diagram of an Item Pool for Geometry



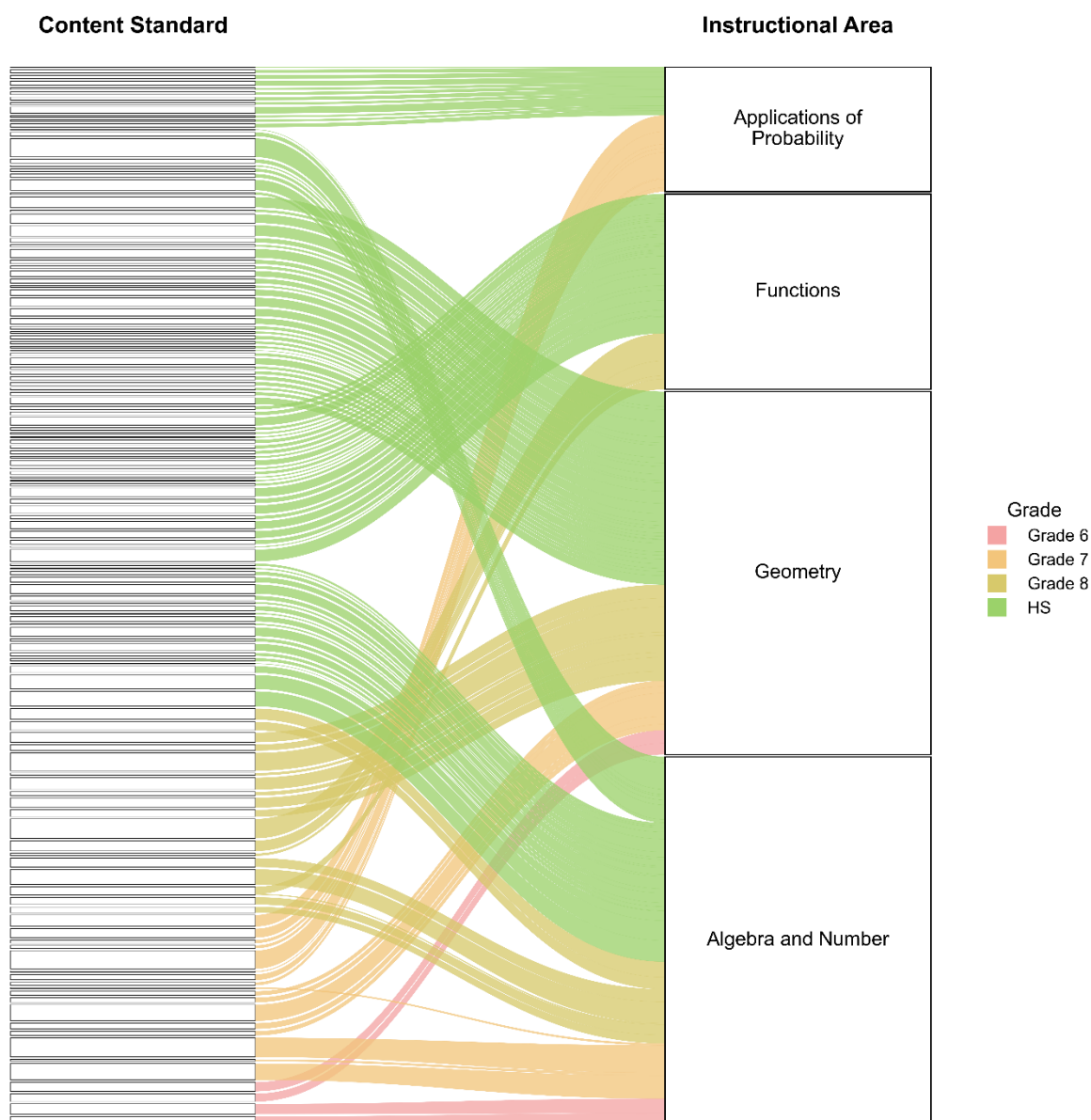
As shown in Figure 3.3, Geometry has a unique structure. Congruence, Similarity, Right Triangles, & Trig has the most items coming mainly from high school. All the prerequisite grades are represented in the instructional area Geometric Measurement and Modeling. Geometric Properties with Equations and Circles is the smallest instructional area, with the least number of items. Grade 6 and grade 8 have a few items in Geometric Properties with Equations and Circles, but the instructional area is predominantly linked to high school content.

Figure 3.4. Alluvial Diagram of an Item Pool for Integrated Math 1



A different structure emerges for Integrated Math 1 (Figure 3.4) than for Algebra 1 (Figure 3.1). The instructional area Algebra and Quantities has the most items, while Geometry has the least. Prerequisite skills come from multiple grade levels, as appropriate, in each instructional area.

Figure 3.5. Alluvial Diagram of an Item Pool for Integrated Math 2



Integrated Math 2 (Figure 3.5) has a similar structure to Integrated Math 1 (Figure 3.4) but with more emphasis on Geometry and less on Functions. The Algebra and Number instructional area has the most items, while Applications of Probability has the least.

Figure 3.6. Alluvial Diagram of an Item Pool for Integrated Math 3

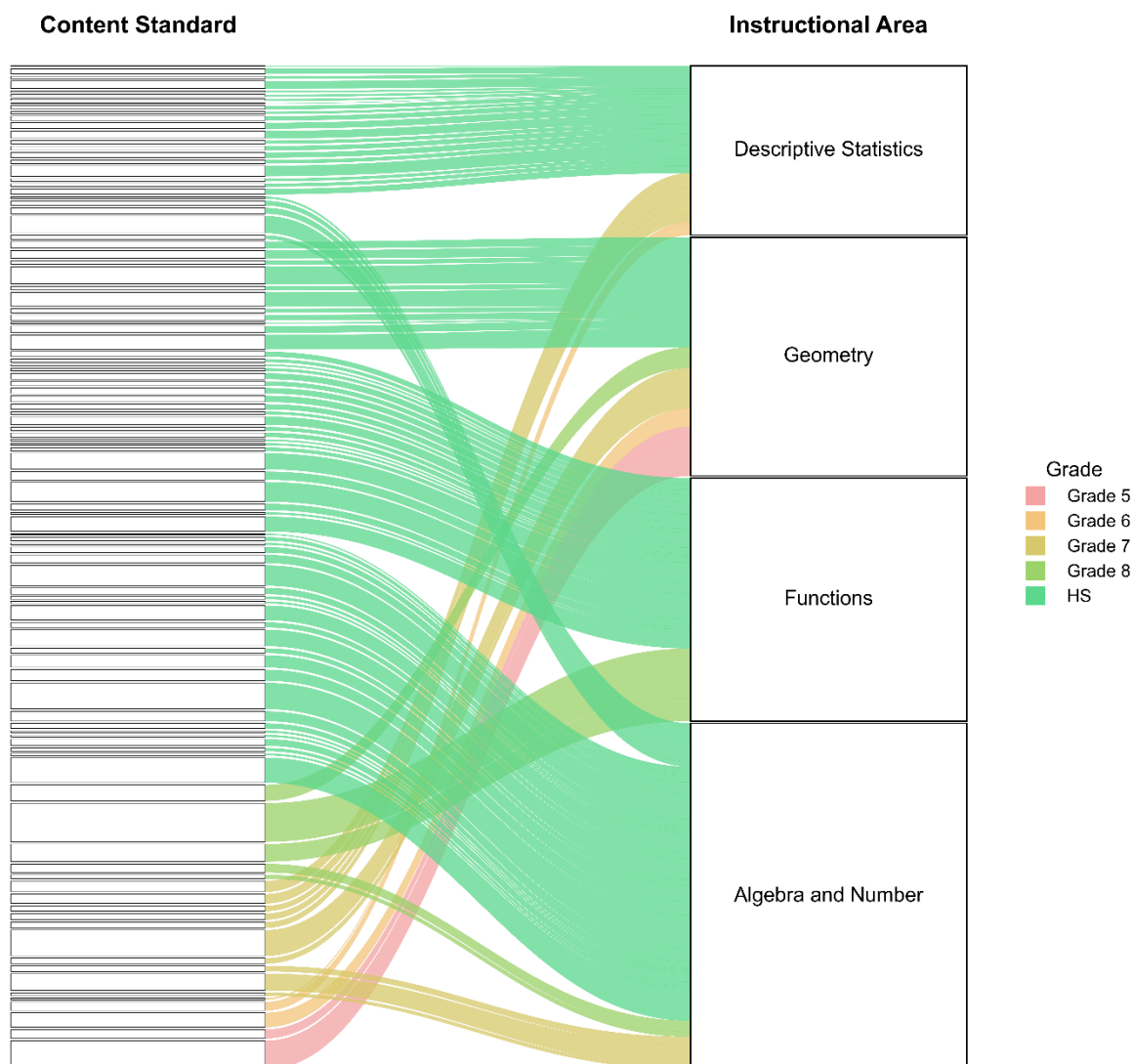
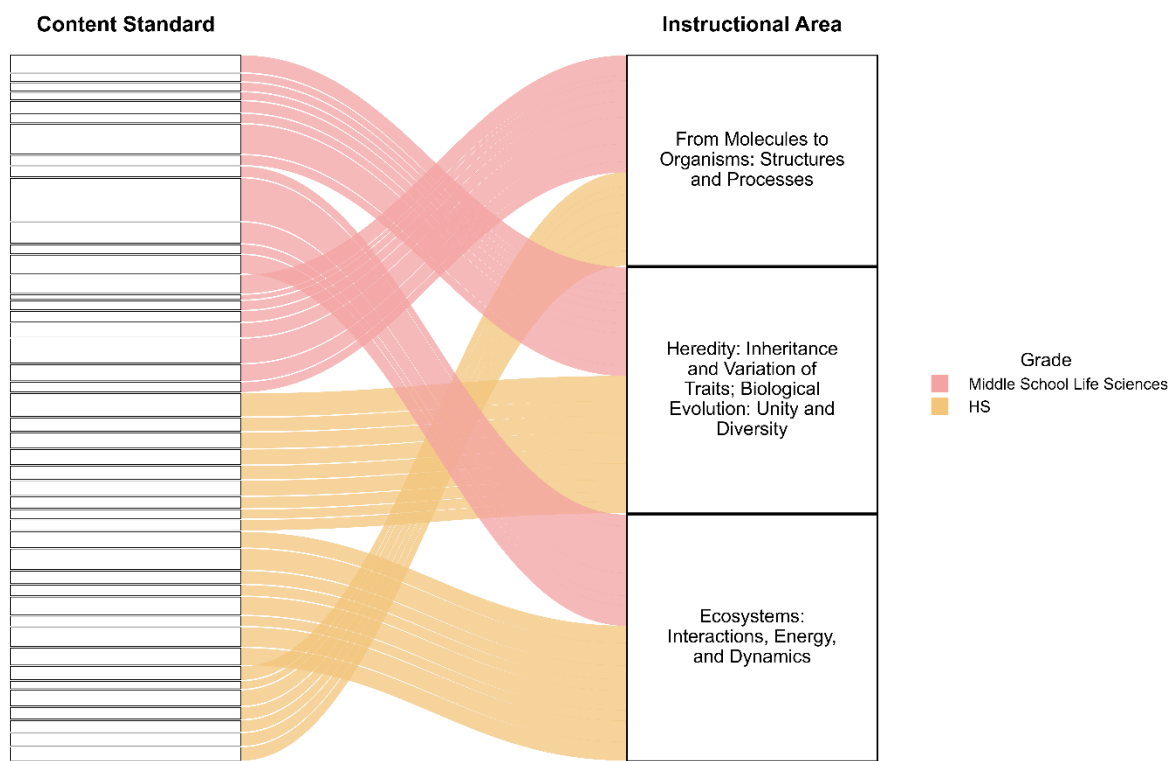


Figure 3.6 shows that Integrated Math 3 has a similar structure to the other two Integrated Math course-specific tests. The instructional area Algebra and Number has the greatest number of items, followed by Functions. Prerequisite standards come from several grade levels and are placed appropriately, with Geometry having standards represented from all prerequisite grades. Geometry contains items beginning at grade 5 that are direct prerequisites for high school Geometry content in order to accommodate lower-performing students.

Figure 3.7. Alluvial Diagram of an Item Pool for Biology/Life Science

7



The Biology/Life Science item pool shown in Figure 3.7 includes relatively similar numbers of items across the three instructional areas. This distribution reflects the comparable emphasis placed on these areas in the content standards. The figure also shows that prerequisite middle school content is represented in each instructional area, supporting measurement across a broader range of student readiness for high school Biology/Life Science content.

The preceding figures pertain to tests developed for a single state and illustrate how that state's content standards are mapped to instructional areas to define a course-specific item pool. Comparable diagrams could be generated for every state using MAP Growth course-specific assessments but are omitted here for conciseness.

3.2. The Item Pool Is Ordered by Grade Level

Using item pools aligned to the Common Core State Standards (CCSS) for course-specific tests in math and to the Next Generation Science Standards (NGSS) for science as examples, the progression of item difficulty over terms and grades provides empirical evidence that the item pools reflect the ordered structure of the target domains and curricula. Within each course-specific pool, item difficulties span a broad range, including prerequisite content and content aligned to the target course. Average item difficulty generally increases across prerequisite grade levels and culminates with items aligned to the course being assessed.

Table 3.1 summarizes item difficulty statistics for the item pools. Similar patterns are observed in other MAP Growth course-specific item pools because many items are shared across pools,

although their alignments to specific state standards may differ. Detailed tables for other item pools are omitted for concision.

Table 3.1. Item Difficulty Summary Statistics for All Items in the CCSS/NGSS Item Pools

Course	Item Grade	N	Mean	SD
Algebra 1	6	761	228.64	18.53
	7	774	233.14	17.36
	8	795	241.88	16.47
	HS	1,115	257.37	17.53
Algebra 2	7	303	238.97	14.06
	8	434	242.81	13.66
	HS	1,437	260.44	18.11
Geometry	6	174	234.33	25.22
	7	447	242.19	22.01
	8	292	250.89	17.05
	HS	844	262.58	20.53
Integrated Math 1	6	706	229.09	19.11
	7	443	234.50	15.21
	8	754	241.19	16.59
	HS	773	253.59	18.31
Integrated Math 2	6	148	234.64	16.15
	7	541	237.86	19.43
	8	693	248.05	15.23
	HS	1,764	260.80	17.74
Integrated Math 3	5	97	238.47	16.15
	6	63	232.63	26.18
	7	240	243.00	16.47
	8	217	245.28	15.92
	HS	1,356	262.13	18.60
Biology/Life Science	6–8	381	216.57	13.11
	HS	781	220.06	13.83

Note. SD = standard deviation. Some items are associated with a grade range; for those items, the summary uses the lower grade in the range.

3.3. Test Blueprints Represent the Content Standards

Blueprints guide how the adaptive test engine constructs an individual test event (i.e., how the test is experienced by an examinee). The example blueprints in Table 3.2 identify the instructional areas and other content features targeted during test administration. In math, these additional content features include Aspects of Rigor (AOR), a framework for cognitive complexity developed by Achieve (2019).

These instructional areas and the other content features are not mutually exclusive; an item may be associated with more than one feature. For example, a Geometry item may be classified as AOR: Application as well as AOR: Procedural. Subscores are reported for instructional areas but not for other content features. The item counts shown in Table 3.2 represent target counts for each content feature within an operational test event. All test events have a maximum of

43 items, including field-test items. Math and science course-specific tests may include up to three field-test items, which are not included in the blueprint counts shown in the table.

Table 3.2. Example Test Blueprints

Test Name	Content Feature	Instructional Area Score	Number of Items
Algebra 1	Equations and Inequalities	Yes	10
	Numerical and Algebraic Expressions	Yes	10
	Functions	Yes	10
	Descriptive Statistics	Yes	10
	AOR: Procedural	No	At least 10
	AOR: Conceptual	No	At least 10
	AOR: Application	No	At least 10
Algebra 2	Equations and Inequalities	Yes	10
	Numerical and Algebraic Expressions	Yes	10
	Functions	Yes	10
	Descriptive Statistics	Yes	10
	AOR: Procedural	No	At least 10
	AOR: Conceptual	No	At least 10
	AOR: Application	No	At least 10
Geometry	Congruence, Similarity, Right Triangles, & Trig	Yes	10
	Geometric Properties with Equations and Circles	Yes	10
	Geometric Measurement and Modeling	Yes	10
	Applications of Probability	Yes	10
	AOR: Procedural	No	At least 10
	AOR: Conceptual	No	At least 10
	AOR: Application	No	At least 10
Integrated Math 1	Algebra and Quantities	Yes	10
	Functions	Yes	10
	Geometry	Yes	10
	Descriptive Statistics	Yes	10
	AOR: Procedural	No	At least 10
	AOR: Conceptual	No	At least 10
	AOR: Application	No	At least 10
Integrated Math 2	Algebra and Number	Yes	10
	Functions	Yes	10
	Geometry	Yes	10
	Applications of Probability	Yes	10
	AOR: Procedural	No	At least 10
	AOR: Conceptual	No	At least 10
	AOR: Application	No	At least 10
Integrated Math 3	Algebra and Number	Yes	10
	Functions	Yes	10
	Geometry	Yes	10

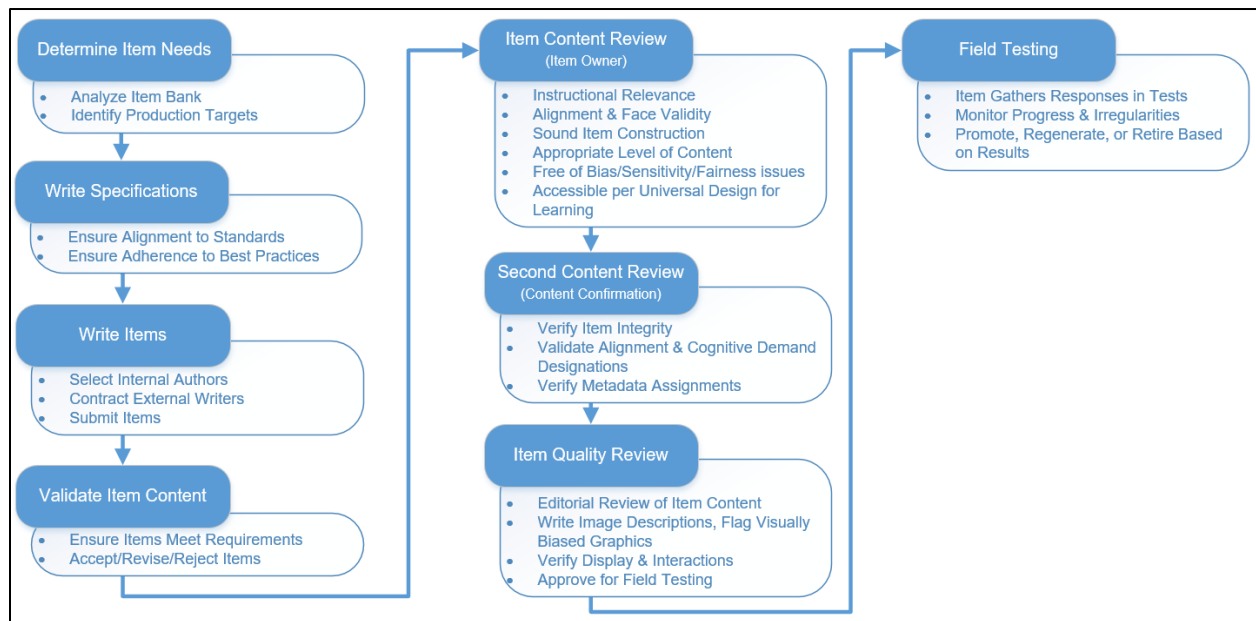
Test Name	Content Feature	Instructional Area Score	Number of Items
	Descriptive Statistics	Yes	10
	AOR: Procedural	No	At least 10
	AOR: Conceptual	No	At least 10
	AOR: Application	No	At least 10
Biology/ Life Science	From Molecules to Organisms: Structures and Processes	Yes	13
	Ecosystems: Interactions, Energy, and Dynamics	Yes	13
	Heredity: Inheritance and Variation of Traits;	Yes	13
	Biological Evolution: Unity and Diversity		

Note. AOR = Aspects of Rigor

4. Item Development Follows a Rigorous Process

MAP Growth assessments utilize an extensive universal item bank with items aligned to state content standards and other content frameworks. Item pools for specific tests are created by selecting appropriate items from this bank. Items are continuously added to the bank through a rigorous process of writing, review, and field testing, as shown in Figure 4.1. Items are also periodically reviewed after they become operational to ensure sustained standards alignment, content accuracy, relevance, fairness, sensitivity, and compliance with style guidelines. Items may be removed due to the results of these periodic reviews, public exposure, or partner feedback.

Figure 4.1. Item Development Flowchart



4.1. Item Specifications

NWEA is committed to creating items that assess what they are intended to assess, adhere to best practices, and are fair and free from bias. NWEA content specialists write items internally or contract out to vendors or freelance content developers. To begin the process, the NWEA content team creates an item acquisition plan based on an item pool analysis and identified areas of need. Once item assignments are given to the content developers, they are provided with ongoing guidance and feedback throughout the development process by NWEA content specialists until items are approved. The NWEA content management system enables content developers to submit items directly into the content review work queues. Writers are provided with guides such as item specifications and the item writing guide as well as ongoing feedback specific to their item-writing assignments.

Item specifications are written to help internal and external content developers create items that are aligned to and assess an intended topic or skill. NWEA item specifications include the following elements of guidance for item writers:

- Describe a direct and demonstrable relationship to areas of need
- Translate an objective into discrete topics and skills
- Ensure that no relevant skills are overlooked when translating an objective

- Focus an item on one topic or skill and specify a grade or grade range
- Identify the cognitive complexity of the topic or skill
- Select appropriate item types for the given topic or skill
- Provide suggestions on the types of answer choice options
- Provide parameters, examples, definitions, and resources when applicable

Content specialists review each specification for clarity, completeness, and alignment to ensure that content developers will understand the types of items expected. The specifications are reviewed and updated on an ongoing basis.

4.1.1. Metadata

During item construction, values for metadata fields are added to each item and reviewed. Item metadata define attributes of the item and provide information for systems to include and exclude items from pools as necessary. Metadata are entered and confirmed by content specialists during each stage of item review. Table 4.1 lists examples of the types of metadata stored for each item.

Table 4.1. Examples of Item Metadata Types

Subject	Item type	Provisional RIT
Grade	Scoring	Operational RIT
Bloom's cognitive level	Allowable tools	Language
Depth of Knowledge	Calculator use	Legal ownership
Aspects of Rigor	Stimulus code	Unit of measure

Item metadata determine how items are selected, administered, and analyzed within an item pool. For example, the subject field ensures that only items from the appropriate content area are selected for a given test. For math and science items, metadata fields for allowable tools (such as rulers or protractors) and calculator type (such as basic or scientific) determine which tools are available to students during testing. Other metadata (including grade, depth of knowledge, and item type) support standards alignment, item-development planning, pool management, and internal psychometric analyses.

When text or graphic assets are associated with an item, content specialists add or confirm element metadata used primarily for internal tracking and analysis purposes. Additional element data are added by the permissions department, including disposition, rights status, copyright information, publisher information, and source documentation. For graphic assets, the asset type, file ID, element location, date, and fulfiller identification information are stored for each asset.

4.2. Item Content Alignment

4.2.1. Alignment to Standards

MAP Growth items are aligned to multiple state content standards and frameworks using a consistent alignment approach that requires a direct and clearly evident connection between the skill or content being measured by the item and the skill or content described by the standard, either as a whole or as a distinct and independent element within the standard. To be aligned, an item must be appropriate for the development level, difficulty level, reading level, and cognitive complexity expected by the standard or targeted element. Items are typically aligned to the most granular level in a standard set's hierarchy rather than to general, overarching standards and are not aligned to grade-agnostic anchor standards.

Items may be aligned to a standard during their initial development—as when an item is specifically created to address a specific standard—or during a review of the intersection of a standard set with the existing item pool. Alignment of existing items to standard sets is facilitated by the extensive and multi-faceted metadata associated with each item and confirmed by content specialists performing eyes-on reviews of individual items to ensure quality. All alignment decisions of new or existing items are confirmed by content specialists. When new standard sets are adopted by educational agencies, content specialists analyze the structure and content as well as any supplemental materials to ensure an in-depth understanding of the standards and consistent application of alignment decisions.

NWEA conducts both internal and third-party alignment studies using a variety of methodologies. These studies provide evidence for MAP Growth alignments across subjects and standard sets, demonstrating alignment at the item-standard level and through the overall test structures. Findings from these alignment studies are reviewed by content specialists and used to inform ongoing alignment practices.

MAP Growth item-standard alignments are maintained in a stable database connected to the NWEA content management system and test-building applications. This system provides an authoritative record of each item’s association with each standard set for a given subject. Alignments may be updated when states release new guidance about standards interpretation or when item content is revised in a way that impacts its connection to a standard. Content specialists also conduct periodic reviews to support the accuracy and consistency of item alignments.

4.2.2. Cognitive Complexity

Webb’s Depth of Knowledge (DOK) and Bloom’s revised taxonomy are two different ways of classifying cognitive expectations and are the most commonly used cognitive expectation classifications in education. To ensure that the MAP Growth assessments include a pool of items that span the full range of cognitive levels and skills, content specialists have created cognitive expectation frameworks that define the target DOK for every standard. The cognitive levels are based on three of Webb’s DOK categories (1997):

1. Recall and Reproduction
2. Skill/Concept
3. Strategic Thinking and Reasoning

Each item in the pool is evaluated and tagged with a DOK level and one of Bloom’s cognitive process dimensions (e.g., remembering, understanding, applying, analyzing) (Anderson & Krathwohl, 2001, pp. 67–68). Additionally, math items have been tagged according to Student Achievement Partners’ Aspects of Rigor (AOR) model (Achieve, 2019).

4.3. Item Content Review

Each MAP Growth item is reviewed by a minimum of three separate professionals (i.e., two content specialists and a copy editor/quality control specialist). An item may be returned to an earlier stage of development or rejected altogether if it does not meet strict review standards. The review process addresses (a) copyright and legal permissions, (b) content accuracy and relevance, (c) cognitive complexity, (d) grammar and editorial style, and (e) accessibility, usability, and interface display.

4.3.1. Copyright and Legal Permissions Review

All items and assets that contain or reference third-party content (including models, diagrams, and data tables developed for MAP Growth Biology/Life Science usage) undergo a comprehensive copyright and permissions review. Each item or asset is evaluated for usability, screened for plagiarism, and examined for permissions-related considerations. HMH's intellectual property compliance department supports NWEA's content development to ensure that no item or asset infringes, violates, or misappropriates any copyright, trademark, trade secret, trade dress, patent, right of publicity, right of privacy, or any other protected rights held by an individual or entity.

Assets are reviewed to confirm the absence of plagiarism, and fact-based items and passages are reviewed to verify factual accuracy.

Public domain materials are validated against original sources, and their public domain status and citations are documented.

Any copyrighted text or visual asset must be formally authorized by the copyright holder. HMH's intellectual property compliance department facilitates and negotiates contractual agreements between HMH and the copyright owner or authorized agent. The department manages the processes for reviewing permissioned assets against contractually defined publishing requirements and for tracking expiration and renewal timelines.

4.3.2. Content Validation: Accuracy and Relevance Review

Concurrently with the copyright and permissions review, items undergo a content accuracy review performed by a content specialist who determines whether the item content meets the requirements outlined in the item specifications and other item development resources. The NWEA content specialist reviews items for the following:

- Content validity
- Instructional relevance
- Contemporariness and currency
- Alignment to standards
- Item construction
- Bias, sensitivity, and fairness

The main purpose of the review is to determine whether a newly submitted item meets basic quality requirements. If the item does not meet the requirements, a content specialist will send the item back to the item writer with a revision request or reject the item so it does not move forward. Items that pass content accuracy and relevance review are moved to the next stage of review.

4.3.3. Item Owner: In-Depth Item Review

A content specialist performs an in-depth review of the item and makes any revisions or edits that are needed. The content specialist uses a checklist such as the one shown in Table 4.2 to validate the item's grade level and standards alignment and assigns a cognitive complexity level to the item by designating Depth of Knowledge, Bloom's, and Aspects of Rigor (for math items) classifications. The content specialist also writes or confirms the equation description for content written in MathML (an application of XML for describing mathematical notations) so that it can be read by a screen reader for math and science items intended for grades 2–12. Finally, the content specialist assigns a preliminary difficulty level (i.e., provisional RIT difficulty value)

needed for field-test purposes. The preliminary difficulty level is based on the observed difficulty of similar items and the content specialist's professional expertise and allows items to be chosen for presentation that closely matches the student's estimated performance level.

A second content review is performed by a different content specialist. This second reviewer attends to the overall editorial and pedagogical integrity of the item and validates the alignment and cognitive demand designations. The content specialist also verifies that the fields have been set appropriately in the NWEA content management system to ensure that the item is ready for field testing, which includes confirming the equation descriptions for MathML as needed.

Table 4.2. Item Review Checklist

Content	Item content is accurate.
NWEA Style	Item writing adheres to the NWEA style guide.
Components	Item includes all required components.
Copy editing	Item uses correct grammar, spelling, punctuation, capitalization, and syntax.
Bias/ Sensitivity/ Fairness	The item is fair for all students. The review ensures: • Content is accessible to all students without a need for prior knowledge. • Item avoids bias (e.g., linguistic, socioeconomic, religious, colorblindness, gender). • Item avoids common issues for ELL students (e.g., idioms, unnecessary phrases). • Item avoids stereotypes. • Item avoids sensitive topics (e.g., smoking, death, crime, violence, profanity, sex).
Item Purpose	Item aligns to standards and is instructionally relevant. The review ensures: • Item aligns to the standard. • Item is instructionally relevant. • Item is not a trick question. • Concept in item is accurately reflected in item resource (passage/graphic). • Item context is appropriate.
Readability	Item and passage readability meets the following criteria: • Item uses an appropriate level of vocabulary and readability for the skill level. • Item includes directions and/or introductory text that is clear, appropriate, and useful.
Passage	A passage meets the following criteria: • Passage is relevant, essential, and engaging. • Passage length is within established guidelines for the intended grade. • Passage citation is correct. • Passage has appropriate permissions for use.
Graphics	Graphics meet the following criteria: • Graphics are accurate, relevant, and clear. • Graphics citation is correct. • Graphics include appropriate labels and titles.
Stem	The item stem meets the following criteria: • Stem is focused, concise, and precise. • Stem uses appropriate terminology, vocabulary, wording, and formatting. • Stem is consistent with answer options.

Answer Options	Distractors and/or the key meet the following criteria: • There is only one key for single-select items. • There is only one correct set of keys for multiselect items. • Key is correctly marked for scoring purposes. • Options are independent (e.g., not overlapping, not logical opposites). • Terminology, vocabulary, wording, and formatting are appropriate. • Options are balanced in length, complexity, and grammatical form. • Distractors are plausible. • Key is not cued. • Options are consistent with what the stem is asking.
Functionality	Item functionality meets the following criteria: • Functionality works as intended. • Number of objects allowed in a container is correct. • Size and type of container are correct. • Item scores correctly and as intended.
Overall Appearance	Item adheres to UDL guidelines.

4.3.4. Item Quality Review

During the item quality assurance review, a copy editor reviews each item as it appears in the system for examinees. The reviewer checks syntax, grammar, usage, spelling, and punctuation, ensuring that the items are formatted according to the comprehensive NWEA Formatting and Style Guide, which addresses technological, visual, and pedagogical considerations, among others. Additional resources used during all reviews to maintain consistency in items are the *Merriam-Webster's Online Dictionary*, *Chicago Manual of Style*, and *Scientific Style and Format: The CSE Manual for Authors, Editors, and Publishers*.

In addition to reviewing item style and formatting, copy editors review each item for visual bias, and image descriptions (alt text) are added to graphics for use by screen readers.⁶ Image descriptions may allow students who use refreshable braille and/or screen readers to answer items that otherwise would be inaccessible. Finally, an editor validates that the item display and interactions are performing as expected in all supported browsers and approves the item for field testing. If at any point changes are required that may impact the content of the item, a content specialist is consulted.

4.4. Field Testing, Calibration, and Psychometric Review

New items are field tested by embedding them in operational MAP Growth test events to reduce the amount of testing time and have students respond to field-test items with as much effort as to an operational item. There is nothing in the test event that indicates to the student whether an item is an operational or a field-test item. The purpose of field testing is to use response data to compute the operational item difficulty on the RIT scale and statistically evaluate the quality of the item.

A provisional difficulty value is used to present the item to students of an appropriate ability level. The system collects data and empirically updates the provisional item difficulty value following every 500 responses. The item remains in field testing and collects data until the quarterly item calibration is conducted by a psychometrician. Items must have at least 1,000

⁶ Image descriptions follow the *NWEA Image Description Guidelines for Assessments*: https://www.nwea.org/uploads/2022/11/Image-Description-Guidelines-for-Assessments_NWEA_2021.pdf

responses to be included in a calibration, although Ingebo (1997) has shown that a sample size of 300 is adequate for accurate item calibrations.

NWEA uses a fixed-person item calibration design. Person parameters for the sample of students responding to an item have ability values estimated from their responses to operational items. The person parameters are fixed to these values when calibrating each item. Item difficulty calibration is done with a proportional curve-fitting algorithm similar to the one used by WINSTEPS (Linacre, 2023). Item review criteria are applied to the calibration results. The review of each item involves a variety of fit statistics and classical item statistics.

Field-test items are calibrated to the RIT scale; each subject (e.g., math, reading, science) has its own scale. The item difficulty parameter is estimated, classical item statistics are calculated, and several fit statistics are calculated. Table 4.3 shows the review criteria for individual field-test items. Items that pass review become operational. All operational items are periodically tested for item parameter drift and content quality after they are made operational.

Table 4.3. Field Test Item Review Criteria

Statistic	Description	Acceptable Values
Non-Convergence	A flag to indicate whether the item calibration converged.	Converged
Extreme	A flag to indicate whether examinees responding to the item all answered it correctly or incorrectly.	Not Extreme
<i>P</i> Value	Classical statistic: the proportion answering correctly	0.05–0.95
Discrimination	Item-theta correlation: shows the relationship between the item and the total test score	0.17–1.0
INFIT	Rasch INFIT statistic	0.5–1.5
OUTFIT	Rasch OUTFIT statistic	0.5–1.5
pexp	Correlation between observed proportion correct and expected proportion correct	0.6–1.0
Distractor	Distractor—total score correlation: shows the relationship between a distractor (incorrect answer) and the total score.	Negative Correlation

Items that did not converge, are extreme items, or fail all review criteria automatically fail calibration and do not become operational. Items that meet all criteria automatically pass calibration and become operational. Items that have converged and are not extreme and have passed some criteria are reviewed by a psychometrician and a content specialist to determine if the item fails or passes. An item that fails after human review may be revised and re-field tested.

4.5. Item Types

Students interact with the assessment in various ways. Interaction types include *selection items* where a student chooses a response from a given set of options, *construction items* that require the student to build a response from the given information, and *generation items* that ask the student to create a response without being provided any response choices. Items can also be presented as *composite items* where responses to multiple interaction types are scored as a single item. All items are scored dichotomously as correct (one point) or incorrect (zero points).

Selection Items include multiple-choice, multiselect, and hot text. For a multiple-choice item, students read the stem and select one and only one response from a list of response options that includes the correct option and multiple distractors (i.e., incorrect options). A multiselect (also known as multiple select) item is similar. Students read the stem and select two or more options from a list of response options. The student must select all correct options and nothing else to earn a point for a correct response. The final type of selection item is a hot text item, also known as a selectable text item. To earn a point for a correct response, students read the item and select a response (e.g., text, equation, symbol) from within a body of text or a table of information. Figure 4.2 through Figure 4.5 show samples of selection items.

Figure 4.2. Multiple-Choice (Science)

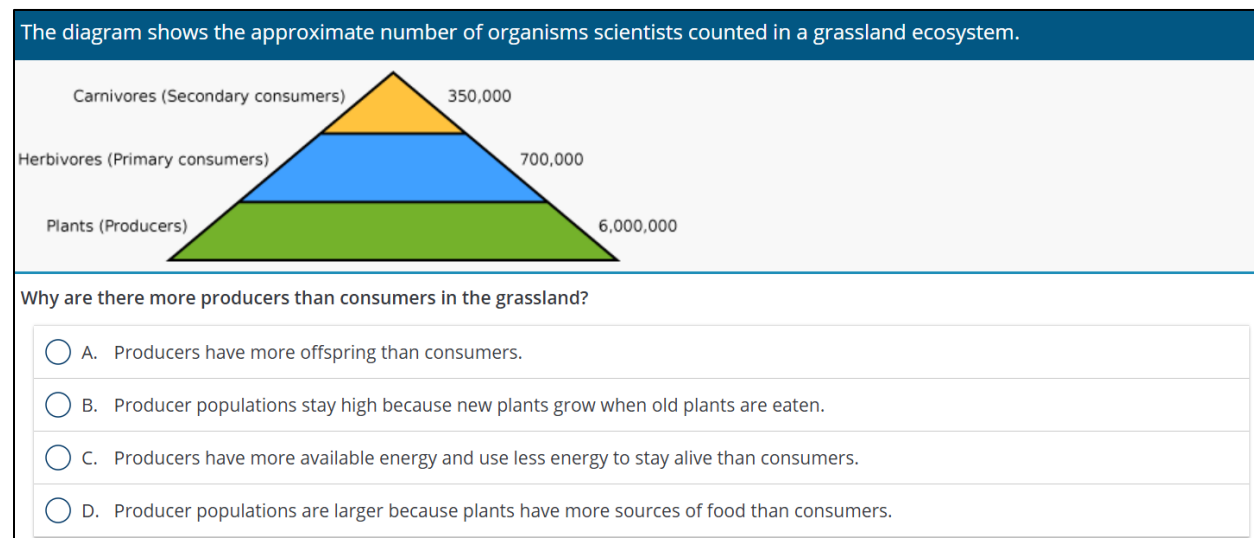


Figure 4.3. Multiple-Choice (Math)

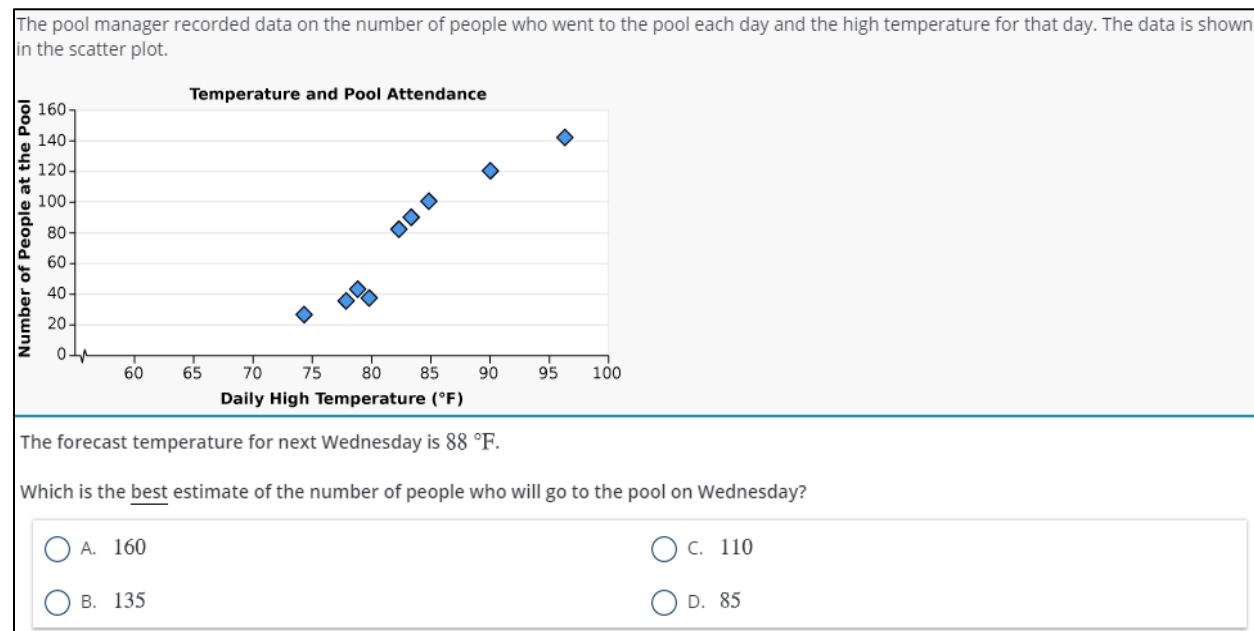
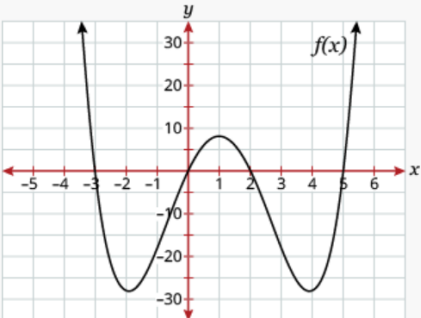


Figure 4.4. Multiselect/Multiple Select (Math)

Use the information to complete the task.

The graph of $f(x) = 0.5x(x + a)(x + b)(x + c)$ is shown.



Choose three numbers that show the possible values of a , b , and c .

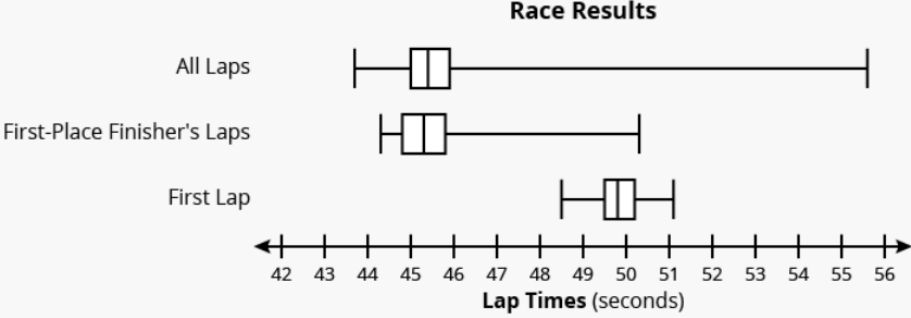
- ☐ A. -5
- ☐ B. -4
- ☐ C. -3
- ☐ D. -2
- ☐ E. -1
- ☐ F. 0
- ☐ G. 1
- ☐ H. 2
- ☐ I. 3
- ☐ J. 4
- ☐ K. 5

Figure 4.5. Hot Text/Selectable Text (Math)

Use the information to complete the task.

An auto-racing event that includes 100 laps around an oval track takes place. The lap time is the number of seconds it takes a driver to complete one lap around the track. The box plots show the lap times of three populations: the whole population, which includes the 100 lap times of each driver in the competition; a sample population that shows the 100 lap times of the first-place finisher; and a sample population that shows the first-lap times for each of the drivers.

Race Results



Choose "True" or "False" for each statement.

Statement	True or False?	
The first-place finisher had the fastest lap time in the race.	True	False
The drivers averaged a slower pace in the first lap than in the rest of the race.	True	False
The first-place finisher's lap times had the least amount of variance of the data sets shown.	True	False

Construction items are defined by drag-and-drop/click-and-click and click-and-pop. These items require students to select an option or options (e.g., words, phrases, symbols, equations) from an area called the toolbar. For a drag-and-drop item, students must move the selection (i.e., drag the selection) to the designated container on the screen. Alternatively, these drag-and-drop

items can interact as click-and-click items where the option is selected (clicked on) and then the designated container is clicked on, which will cause the option to appear in the container. Using click-and-click functionality makes these items compatible with keyboard navigation.

For a click-and-pop item, the selection automatically “pops” to the container on screen. All parts must be selected and placed into the right container for a student to earn credit. A student builds their response to the question by selecting among available options. Figure 4.6 and Figure 4.7 show examples of construction items.

Figure 4.6. Drag-and-Drop/Click-and-Click (Science)

Use the information to answer the question.

Students test how quickly they can hit a button after hearing a sound. The student with the quickest time took 0.17 seconds. The students wonder why no one was faster than 0.17 seconds. They make a model to show how the nervous system processes the sound.

Why does the nervous system take time to respond to hearing a sound? Move one statement into each box to complete the model. Statements may be used more than once or not at all.

Input

Process

Output

Muscles
receive
signals.

Signals are
processed in
the brain.

Signals are
processed
in the
spinal cord.

Finger
pushes
button.

Signals
travel along
nerve cells.

Sense
receptors in
ears respond
to sound.

Figure 4.7. Click-and-Pop (Math)

Use the graph to complete the task.

Choose all the factors of the polynomial shown in the graph.

x
 $2x$
 x^2
 x^3
 $(x^2 + 4)$
 $(x^2 - 4)$
 $(x + 2)^2$
 $(x - 2)^2$

Generation items do not present students with options. Instead, they are text-entry items that require students to use a keyboard to type a response to a question. Correct responses, and known variations of them, are scored as correct. Figure 4.8 shows an example of a text-entry item.

Figure 4.8. Text Entry (Math)

Use the equation to answer the question.
$\sqrt{x+9} = x-3$
What extraneous solution is introduced if the equation is solved by first squaring both sides of the equation? Enter the answer in the box.
$x = $

For an item set, students are presented with a set of items that all focus on a single reading passage or a narrowly defined topic (also known as a stimulus). However, item sets are not currently in use in course-specific tests.

Items on course-specific tests may also use a composite format where multiple interactions are presented and scored as a single item. Figure 4.9 and Figure 4.10 show examples of composite items.

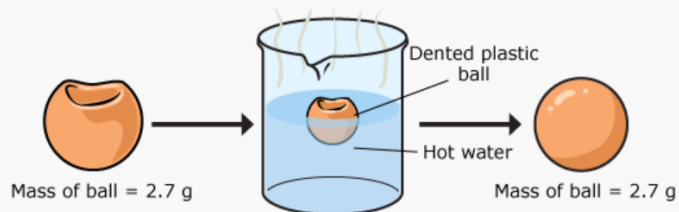
Figure 4.9. Composite Item (Math)

This question has two parts. Use the information to answer Part A and Part B.
When Sonia buys a new car, she is told she will be able to drive 30 miles for every gallon of gas while driving on the highway. She wants to know how many miles she can drive for every gallon of gas while driving on city roads.
She drives 174 miles on the highway and 236 miles on city roads. She uses a total of 16 gallons of gas.
Part A Which equation can be used to find the number of miles, x , Sonia can drive for every gallon of gas while driving on city roads?
☐ A. $(30)(174) + 236x = 16$
☐ B. $(30)(174) - 236x = 16$
☐ C. $\frac{30}{174} + \frac{x}{236} = 16$
☐ D. $\frac{174}{30} + \frac{236}{x} = 16$
Part B How many miles can Sonia drive for every gallon of gas while driving on city roads? Round the answer to the nearest tenth. Enter the answer in the box.
miles per gallon

Figure 4.10. Composite Item (Science)

This question has two parts. Use the information to answer Part A and Part B.

A student wants to remove a dent from a hollow plastic ball used for table tennis. He reads that table tennis balls are filled with oxygen gas. He decides to put the dented ball into hot water to see what happens. The diagram shows the results.



Part A

Which statement explains the results of the investigation? Choose one explanation.

- ☐ A. Oxygen molecules inside the ball move farther apart and push out the dent.
- ☐ B. Oxygen molecules inside the ball fill with heat, grow larger, and push out the dent.
- ☐ C. Hot air molecules enter the ball. The increased number of molecules pushes out the dent.
- ☐ D. Hot water molecules enter the ball. The increased number of molecules pushes out the dent.

Part B

Which information is evidence that supports this explanation? Choose all the supporting evidence.

- ☐ A. Ball loses its dent.
- ☐ B. Volume of the ball increases.
- ☐ C. Mass of the ball stays the same.
- ☐ D. Ball floats on the surface of the water.

5. Test Administration Is Standardized and Secure

MAP Growth assessments are fully adaptive, and each student experiences a unique test based on their responses to each item. MAP Growth course-specific assessments are untimed and typically take approximately 45–60 minutes per content area. MAP Growth can be administered up to four times a year (fall, winter, spring, and an optional fourth administration in summer). A MAP Growth administration requires a proctor computer that allows the proctor to monitor and control the student testing, as well as student devices with lockdown browsers. There are three main steps to testing:

1. Proctor creates a testing session.
2. Students sign in and join the testing session created by the proctor.
3. Proctor supervises students and assists them with actions such as pausing and resuming their test if needed.

The NWEA test delivery platform supports nearly 77 million test events each year. The platform has delivered uninterrupted service with 525,000 students actively testing (i.e., “concurrent” users). The most recent configuration has been certified and tested for at least 650,000 concurrent users.

Both in-school and remote test administration are possible. Each is handled by a trained test proctor who is available to assist students with technical challenges and other issues that may arise during testing. During in-school administrations, the proctor is in the room with students; during remote administrations, proctoring is facilitated through a video meeting, chat, email, or other communication channel. Non-adaptive practice tests are available online to familiarize students with the types of questions and item types used in MAP Growth.

5.1. Test Engagement Functionality

When students are motivated and engaged during testing, their results are more likely to accurately reflect what they know and can do. In 2017, NWEA introduced test engagement functionality that detects rapid guessing in real time and alerts proctors so they can re-engage students. A summary of this functionality is as follows:

- Students receive a message at the start of the test encouraging them to remain engaged.
- When rapid guessing is detected, the proctor is notified and the test auto-pauses so the proctor can re-engage the student and resume the test.

MAP Growth uses response-time information to help identify rapid guessing. The average time students take to answer each unique item is used to determine whether a response was likely made without full engagement. After a student rapid-guesses on one item, the difficulty of the next item is held at the same level to help prevent downward drift in the adaptive algorithm. After three consecutive rapid guesses, the proctor is notified and the student’s test is paused so the proctor can intervene and re-engage the student.

Reporting for the test event includes the percentage of the assessment during which the student rapid-guessed and the estimated impact that disengagement may have had on the student’s overall RIT score.

5.2. Administration Training

Test administration training is provided online. Through NWEA Professional Learning Online, staff members have access to a series of online training courses that will prepare them to administer MAP Growth assessments and utilize student data. The online format allows flexible scheduling and pacing as well as the opportunity to review content as needed. More information can be found at <https://www.nwea.org/professional-learning/map-growth-professional-learning/map-growth-basics/>.

5.3. Practice Tests

Practice tests are available online at <http://practice.mapnwea.org> for students to familiarize themselves with the assessment. They provide the same access and functionality as the real MAP Growth tests. Students are encouraged to use the embedded universal tools or a designated feature or accommodation, if needed. To take the practice tests, users must enter a generic username and a password. For MAP Growth tests, the username and password are both “grow.” Details about practice tests are as follows:

- Not adaptive
- Not scored
- No proctor control
- Available through any supported browser and any supported device
- Available for multiple grades and content areas
- About five items in length, depending on the grade

The purpose of the practice tests is to introduce students to the test’s form, function, and item types, not to prepare them for the content. Students preparing for course-specific assessments should select any grade from 6 through 12 and then choose the Math K–12 or Science 3–8 practice test to see the item types and tools available for the math and science tests.

5.4. Test Security

Inadequate security procedures pose a risk to assessment systems. Violations of test security may compromise the integrity of results and call into question the trustworthiness of score interpretations. A common concern for adaptive tests is whether item pools are large enough to prevent content from being memorized, compiled, or shared by groups of students or educators. Well-designed adaptive tests such as MAP Growth draw from large item pools and include several features that help protect test and item security.

MAP Growth systems leverage several inherent security advantages. First, a group of students within a classroom or computer lab is likely to view hundreds of different items in any single administration of the test, making it unlikely that students will see the same content at the same time or see items used as examples in a classroom. Second, once a student has viewed an item, they will not see that item again for a specified period of time. Third, the size of the item pool makes it impossible to anticipate and practice for the specific questions that will appear during a test event. Finally, if there are concerns about the integrity of a score, a student can be retested with a new set of items.

Other security features are built into the test administration itself. When a student logs into a session, the test is not started and no test items are visible to the student until the proctor has confirmed the student and activated the test session through the proctor dashboard. Furthermore, item responses are not stored or cached locally; instead, responses are captured in real time and stored on secure servers before presenting the next item to the student. In

addition, a lockdown browser can prevent students from initiating other browser sessions or accessing other content on the testing device unless they exit the test.

The processes and tools summarized in Table 5.1 **Error! Reference source not found.** further support test security and help protect the integrity of test administration, providing educators and students with a reliable user experience.

Table 5.1. Test Security Before and During Testing

Before test administration	• Student and educator data are rostered through secure system applications. • Only specific user roles, approved and authorized within the district and school, can log into the system to access test administration features. • All testing devices are prepared by installing the secure testing browser/app.
During test administration	• Only approved and authorized proctor roles can start the test by providing a secure test session key for all students in the testing lab/classroom. The proctor has the control to start, pause, and resume testing for all students in the classroom or individual students if necessary. • Student test taking is only possible with a secure testing browser. • There is a district configuration that can be set to prevent retesting. • If students require any testing accommodations, such as TTS, proctors can assign those specific accommodations to students based on their IEP/504 needs and ensure appropriate device setup for those tests (e.g., earphones for TTS). • Student test taking is only allowed during the testing window. All tests are closed and access is removed upon closing of testing window.

5.4.1. Data Security

All MAP Growth data transmissions (i.e., testing and response data) are encrypted and secured using TLS 1.2 AES 256 encryption methods. Test data is stored in highly secure Tier 3 data centers located in the continental U.S., operating with redundant power, internet, and backup systems powered by diesel generators. All servers, disk storage, and network infrastructure within each data center are redundant, protecting against unavailability due to a single hardware failure. NWEA operates two geographically disparate data centers with data replication for failover if one data center becomes inoperable. Personally identifiable student information is encrypted at rest in the systems. More information on NWEA Information Security can be found at <https://legal.nwea.org/map-growth-information-security-whitepaper.html>.

5.4.2. Role-Based Access

Access management is a critical component of test security. MAP Growth uses role-based access controls that allow partners to assign users only the permission needed to perform their responsibilities, as outlined in Table 5.2. This approach supports separation of duties and helps partners control which actions users can take and which data they can access. When partners plan their access-control strategy, MAP Growth supports the principle of least privilege. Each role has specific permissions that govern access to implementation, configuration, data management, testing, and reporting tasks. Each user has a unique username, and one or more roles can be assigned to that account. Only designated roles can create or modify student profiles, which helps limit unauthorized or inappropriate changes to student information. More information on MAP Growth roles and responsibilities can be found at https://teach.mapnwea.org/impl/QRM2_Roles_and_Responsibilities_QuickRef.pdf.

Table 5.2. Access Roles, Permissions, and Responsibilities

Role	Permissions & Responsibilities
System Administrator	• Assign MAP Growth roles for any user, including themselves • Add or edit users in MAP Growth and reset user passwords • Modify MAP Growth preferences for the organization • Mark the test window complete
District Assessment Coordinator	• Assign MAP Growth roles for any user except System Administrator • View operational reports • Add or edit users in MAP Growth and reset user passwords • Modify MAP Growth preferences for the organization • Mark the test window complete
Data Administrator	• Assign MAP Growth roles for any user except System Administrator or District Assessment Coordinator • View operational reports • Add or edit users in MAP Growth and reset user passwords • Add or edit students • Import student/staff roster • Add or edit students in MAP Growth, including permission to merge students and exclude or assign test events
District Proctor	• Proctor any students within the district • Set up and conduct student testing • Add or edit students in MAP Growth
Administrator	• Limited to assigned schools, will likely be a school principal or vice principal • View student and class reports • View reports for the school
School Assessment Coordinator	• Limited to assigned school(s) • Edit students in MAP Growth
School Proctor	• Proctor any students in assigned school(s) • Set up and conduct student testing
Interventionist	• Limited to assigned schools, will likely be a special education teacher or similar role • View students within their school and add them to custom groups for instruction and reporting

6. Test Events Sample the Domain and Adapt in a Principled Way

All test events are designed to satisfy both statistical and content requirements. In a linear test, balancing these two aspects is straightforward because a fixed set of items is selected in advance and administered to all students. In a computer adaptive test (CAT), test design is more complicated because each test event must not only meet statistical and content requirements but also tailor itself to an individual examinee.

A sequential item-selection algorithm dynamically creates a test as it goes by choosing items one at a time according to an examinee's responses to previously administered items. At each step, the algorithm must determine which item best supports the intended statistical precision and content coverage. Psychometricians have developed a variety of item-selection algorithms for balancing these goals, each with its own strengths and limitations, though all are intended to satisfy the requirements established in the test design.

This chapter provides a general overview of CAT algorithms, followed by details about the MAP Growth CAT engine. Evidence is then presented to show that the MAP Growth engine functions as designed and delivers test events according to test blueprints.

6.1. Computer Adaptive Test Administration

6.1.1. Statistical Requirements

Statistical requirements are derived from the underlying item-response model and address the measurement precision and reliability of a CAT. Maximum information item selection is a procedure where the most informative item is selected at each step of the test. The ultimate result is a test that yields maximum information (i.e., maximum precision) about an examinee's ability (test score) for the given set of administered items.

The goal of maximum information item selection is to create a test with the largest possible test information for a given number of items. This can be achieved by either selecting the item with the largest item information value at each step or by randomly selecting an item from a group of items that all have the largest item information. Given that item selection during the test is based on a student's momentary ability estimate, the estimate at the beginning of a test may differ from the final ability estimate. Thus, items selected early may not be optimal for the final ability estimate. This limitation can be overcome with a termination rule whereby the test continues to administer items until the SEM is reduced to an acceptable value. While the most optimal test may not be achieved for every examinee, a key benefit of adaptive testing over fixed-form testing is that each examinee's SEM will never exceed the value specified in the termination rule.

The maximum information approach to item selection only accounts for measurement precision. It is simply a function of person ability and item difficulty, with no consideration of test content beyond its relationship to item difficulty. Therefore, to meet test-specification requirements, the item selection algorithm must be redefined to explicitly account for content requirements.

6.1.2. Content Requirements

Content requirements are defined by test blueprints. They are not part of the item-response model and cannot be controlled by an algorithm that solely relies on maximum information item selection. To produce a test event that satisfies content requirements, either maximum information item selection must be augmented or the optimization problem must be redefined.

Heuristic item-selection methods use either item pool partitioning or a mathematical function to select items and meet statistical and content requirements. However, a limitation of heuristic methods is that an optimal test is not guaranteed. The greedy nature of item selection picks an item that is the best at the moment but may not be the best item for a complete test. Heuristic methods may not be mathematically optimal, but they are fast, considerably easier to implement in practice, and the calculations are much simpler.

Constrained CAT (C-CAT; Kingsbury & Zara, 1989) is an algorithm that works by partitioning the item pool into mutually exclusive content categories, identifying a category that has not reached its target number of items, and then selecting the most informative item from that category. Randomesque exposure control can be added by randomly selecting the most informative items from a group (Kingsbury & Zara, 1989) within the category. MAP Growth tests have used C-CAT since the assessments were first offered in computer adaptive format. C-CAT is effective when only a small number of mutually exclusive content categories are used. However, when items have multiple and overlapping content classifications, partitioning the item pool into groups can create empty or very sparse categories, resulting in over exposure or item starvation (in which no suitable items are available).

Heuristic methods based on mathematical functions of statistical and content requirements can more easily accommodate items with multiple content categories. Three examples are the weighted deviations model, the weighted penalty model, and the maximum priority index.

In Stocking and Swanson's (1993) weighted deviations model (WDM), the difference between observed item counts and the lower and upper bounds of the target count for each content category is calculated and summed across all content categories and combined with item information. The sum may be weighted, with weights indicating the importance of each content category, and the item with the lowest weighted deviation is selected. A key advantage of the WDM is that content requirements do not need to be mutually exclusive. Moreover, the WDM can explicitly indicate that one item cannot be administered if another item has already been given. That is, suppose one item gives away the answer to another item, the exclusion requirement indicates that the two items may not appear on the test at the same time (i.e., they are "enemy" items). Another advantage of the WDM, and all other heuristic models, is that there is always a solution to the objective function because it reframes mathematical constraints into preferences. An item will always be available for selection as long as the item pool is not exhausted. The tradeoff to using preferences instead of constraints is that the target content requirements may not be exactly met under some conditions.

In the weighted penalty model (WPM; Shin et al., 2009), a total penalty function is calculated as a weighted sum of a statistical penalty (i.e., the negative of the normalized item information) and a composite content penalty. The content part of this total is calculated in several steps. First, a penalty for each individual content requirement is calculated. Next, the maximum value of all penalties is identified, and then all content penalties are divided by the maximum penalty to normalize them. Finally, a composite content penalty is calculated as a weighted sum of the individual normalized content penalty values. After the total penalty function is calculated, the item with the lowest total penalty is selected. Exposure control can be incorporated into the total penalty, or it can be implemented as a Randomesque procedure by selecting among a group of items with the lowest penalty value.

The WPM is effective at meeting statistical and content requirements (He et al., 2014). However, it has several practical drawbacks. Although the WPM provides great flexibility by allowing each part of the penalty function to have a weight, the choice of weights is arbitrary. In

practice, extensive trial and error may be needed to identify appropriate weights and “fine-tune” its performance (He et al., 2014). Another disadvantage of the WPM is that it requires two passes over the entire item pool, once to calculate each individual content penalty and again to normalize each individual penalty. Two passes over the entire item pool for each item selection may notably slow item selection to a point that it adversely affects an examinee’s test experience. Finally, an additional step of “color groups” is needed to prevent one constraint from interacting with another and causing it to exceed its upper bound. This extra step is not factored into the total penalty function; it sits outside of it. As such, the total penalty function does not capture every aspect of selecting an item.

The maximum priority index (MPI; Cheng & Chang, 2009) is a heuristic method that is computationally much simpler than the WPM yet is quite effective at delivering a CAT that includes many content requirements. It involves the calculation of a priority index for each item in the pool and then choosing the item with the highest priority index. The priority index not only quantifies the statistical and content aspects of an item but may also include a function for exposure control. Heavily used items in the pool could be deprioritized directly through the index. Randomesque exposure control is also possible by selecting from among the group of items with the highest priority index.

A notable difference between the MPI and other heuristic methods is that the priority index is a continued product instead of a sum. As a result, it is a conjunctive heuristic model instead of a compensatory one. If any priority function is zero, then the entire index is zero. This feature helps ensure that no target is exceeded, but it could make some targets difficult to achieve, especially if the targets concern content features that overlap with other content features. He et al. (2014) also noted that the zero-valued index could also result in random item selection or item starvation when the index is zero among all available items.

6.1.3. Enhanced Item Selection Algorithm

In addition to the limitation of being undefined when the target value is reached, if any content feature has reached its target, the function will be zero and, ultimately, the MPI itself will be zero. This conjunctive aspect of the MPI may be desirable for some uses, such as mutually exclusive content features, but another limitation is that some features may never reach their targets if they are paired with other features that have.

NWEA’s Enhanced Item Selection Algorithm (EISA) overcomes these two limitations of the MPI. It is a compensatory method defined in such a way that items can be selected by the item information alone if all of the relevant content features have met their targets. This new method also expounds on the notion of a content function. It describes various reward functions that prioritize item selection according to qualitative and/or quantitative item characteristics or examinee attributes.

The approach NWEA has taken with EISA is to assign a reward to each content feature. Reward functions, f_k , have a maximum value of 1 that decreases to 0 as the feature becomes less relevant in order to satisfying the test model. Some rewards also impose a cost (a negative reward) as the target value of the feature is exceeded. After calculating the reward for each individual feature, a total reward, R_j , is computed for each item, as shown in Equation 3.

$$R_j = \lambda I'_j(\theta) + (1 - \lambda) \left[\frac{\sum_{k=1}^K c_{jk}(w_k f_k)}{K} \right] \quad (3)$$

Here, notation was defined previously, and $I'_j(\theta)$ is the normalized item information function shown in Equation 4.

$$I'_j(\theta) = \frac{I_j(\theta)}{\max\{I_j(\theta)\}} \quad (4)$$

The information weight, λ , and the content weight, $(1 - \lambda)$, sum to unity. As such, the total reward is a weighted average of functions representing statistical and content requirements. Because item information is normalized and content feature reward functions have a maximum value of 1, the maximum value of the total reward function is also 1.

The total reward is calculated for each item. Items are then listed in descending order of total reward (i.e., reward listed), and the item with the largest value is selected. If multiple items tie for the largest value, then one is selected randomly. Having tied values is ideal from the standpoint of exposure control because the random selection of an item results in a randomesque exposure-control mechanism (Kingsbury & Zara, 1989).

In contrast to the MPI, the compensatory nature of the EISA model uses a sum of content feature functions instead of a product. The advantage is that the equation is still defined when a feature has met its target value, and other features retain their priority. Moreover, the index reduces to a function of item information if all features have met their targets. Information drives item selection when all content requirements have been satisfied.

The specific content reward functions, f_k , used in EISA are proprietary and are not shared publicly. There are functions that represent each instructional area and, in math, an item's cognitive complexity, as well as a function that represents the relationship between a student's grade and the grade associated with the standard to which the item is aligned. Taken together, the adaptive engine uses an examinee's grade level and ability estimate to select items based on difficulty, content, and grade-level alignment.

6.1.4. MAP Growth Item Selection

MAP Growth's traditional item selection algorithm took item difficulty into account and adapted by finding an item with a difficulty equal to, or close to, a student's ability at that point in the test. If the student answered the item correctly, the student's momentary ability estimate would increase, leading to the selection of a more-difficult item. If the student answered incorrectly, the student's momentary ability estimate would decrease, and an easier item would be selected. Its adaptive item selection also balanced the number of items per instructional area so that each area had a similar number of items by the end of the test. Besides the instructional area, no other content factors were involved in this original adaptive item selection algorithm. Students could see items aligned to standards from a range of grade levels above and below their own as long as the item difficulty was suitable. The traditional algorithm progressed through multiple phases, where each phase emphasized different aspects of item information and content. It also included item exposure control and longitudinal constraints that prevented the overuse of items and displaying the same item to a student more than once in a given period of time.

In 2023, NWEA introduced its Enhanced Item Selection Algorithm (EISA). A key feature of this algorithm is the way it incorporates the student's grade into its adaptive item selection. Using the item and student grade in selection facilitates the ordering of content on an adaptive test beyond ordering by item difficulty alone. The selection of an item is weighted by its proximity to the student's grade. An item aligned to a standard with a grade level that equals the student's

enrolled grade has more weight (i.e., greater priority or reward) in item selection than an item one grade above or below the student's enrolled grade. For course-specific tests, a student's enrolled grade is based on the grade associated with the course and not the student's actual rostered grade. The weight decreases as the distance between the item's standard grade and the student's grade increases. For example, items aligned to a high school Biology standard are more likely to be shown to a student taking the high school Biology/Life Science test than items aligned to an eighth-grade Biology standard. Similarly, while items aligned to an eighth-grade math standard are less likely to be shown to a student taking the Algebra 1 test than items already aligned to an Algebra 1 standard, they are more likely to be shown than items aligned to seventh-grade standards, and so on. The priority of selecting an item is highest when the item's standard grade matches the student's grade; the priority decreases as the item's standard grade differs from the student's grade. In course-specific tests, the priority is always highest for items aligned to standards for that course because all students taking the test are treated as being enrolled in that course and its associated grade. The EISA continues to include item-exposure control and longitudinal constraints to prevent item overuse and limit repeated exposure to the same items over time.

The EISA uses item difficulty in item selection in the same way as the traditional algorithm. Items with a difficulty equal to a student's ability are the most informative and are more likely to be selected than other items. A key difference between the traditional algorithm and EISA is the way that the grade-level priority interacts with the effect of item difficulty in item selection. In course-specific tests, the algorithm adapts off grade depending on the student's ability level. Students will see below-grade items when those items are more suitable than on-grade items. A low-performing student is likely to see more prerequisite below-grade items than a typical student, and a high-performing student is likely to see more-difficult items than a typical student. The reason is that the difficulty of the on-grade item pool may be too high for a very low-performing student, and a prerequisite below-grade item will provide more information about that student's ability level.

The EISA is intended to select items in a way that resembles the type of curriculum and instruction experienced by a student. Typical students will experience mostly on-grade or on-course instruction; therefore, the vast majority of items a typical student will see on a test event will be on-grade items. However, low- and high-performing students are more likely to experience off-grade instruction than a typical student. A low-performing student may be given additional teaching in prerequisite skills or participate in supplemental instructional programs that focus on knowledge, skills, and abilities to allow the student to access on-grade or on-course content. The goal of the instruction is to rapidly move students to achieve end-of-grade standards. A high-performing student may experience enriched instruction that goes beyond typical on-grade instruction as part of a gifted and talented program or move more quickly or explore in greater depth the content of a given course. The curriculum experienced by students may include a mix of prerequisite and course-level content. The EISA reflects this instructional reality by prioritizing course-aligned items while allowing prerequisite items to be selected when needed in order to better match a student's estimated ability level.

For course-specific tests, item selection outside the target course should follow a specific pattern. Low-performing students may receive more prerequisite items than other students, while high-performing students should see more difficult items aligned to the target course. When prerequisite items are selected, items aligned to content closest to the target course should be selected more often than items aligned to more distal prerequisite content. The lower the student's estimated ability level, the more likely the test is to include prerequisite items from earlier points in the content progression. Conversely, students with higher estimated

achievement should receive more-challenging course-aligned items rather than items from content beyond the target course.

6.2. Simulation Procedures and Test Publishing

With a fixed-form test, every test form is designed to meet the test blueprint. It is known in advance of testing if the forms adequately align with standards and meet other criteria specified in the blueprint. A CAT is different because each student receives a test that is customized to their ability level, with that customization occurring dynamically. The test is tailored to student performance as the student is completing the test. It is not known before administration if every test event will satisfy the blueprint. Therefore, the quality of a CAT depends on the breadth and depth of the item pool. A pool that covers all content requirements is one that has breadth; a pool that has enough items to span the range of difficulty needed for a given population of students while satisfying content requirements is a pool that has depth.

Simulation studies are conducted before operational administration to evaluate whether the item pool is sufficient and whether the adaptive engine can deliver test events that meet blueprint requirements. In a simulation study, “simulees” are generated to represent the target testing population. Each simulee is assigned a known ability value drawn from a distribution with the same mean and standard deviation as the target population. A CAT is then administered to each simulee. A simulee’s response to an item is generated by using the true ability score and known item difficulty parameter in an item response model (e.g., the Rasch model) to calculate the probability of a correct response. The probability is compared with a random draw from a uniform distribution with values ranging from 0 to 1. If the probability is greater than or equal to the uniform random draw, a correct response is recorded. Otherwise, an incorrect response is recorded. The simulee’s momentary ability estimate is updated after each response and used by the CAT algorithm to select the next item, just as it would with a real examinee. The process continues until the test is completed. Each simulee’s final score is calculated using the generated responses. The final calculated score is then compared with the true ability score to evaluate score recovery (how close an estimated score reflects the true score). Additional statistics are calculated to evaluate other characteristics of the CAT.

Conducting a simulation study prior to the operational administration serves two primary purposes: to determine whether the item pool can support feasible item selection for students across the achievement distribution and to evaluate whether the item-selection algorithm produces test events that satisfy blueprint requirements to an acceptable degree. Simulation results provide evidence that students receive comparable content representations and that scores have the same intended meaning across different adaptive test events.

Table 6.1 lists the aspects of the simulation results that are evaluated and summarized in a report. Each report is thoroughly reviewed against 14 or more predefined criteria, and an overall decision of Pass, Qualified Pass, or Fail is assigned. A test is published if it results in a Pass or Qualified Pass decision. For a Qualified Pass, a plan is developed to enhance the item pool and address limitations identified in the simulation. A test that receives a Fail decision is modified, such as by adding items to the pool or revising the test model, and then submitted for another simulation.

The simulation criteria provide evidence about multiple aspects of test quality and operational readiness. Content criteria evaluate whether the test contains the correct number and type of operational and field-test items and whether required content areas are adequately represented. Score estimation criteria assess the accuracy of score recovery using metrics such as root

mean squared error (RMSE) and the correlation between true and estimated scores to confirm reliability. Measurement error criteria evaluate the precision of scores, focusing on standard error of measurement (SEM) thresholds for total scores and subscores, both overall and by student decile, to guarantee consistency and fairness. Adaptivity criteria examine how well item difficulties align with student abilities, ensuring the test adapts appropriately to individual performance levels. Item exposure criteria monitor the frequency with which items are presented, aiming to prevent overexposure and maintain test security. Each criterion is assigned a status, such as Pass, Flag, or Not Applicable, and combined results are used to evaluate the psychometric quality and operational readiness of the test.

Table 6.1. Simulation Study Evaluation Points

Category	Evaluation Point
Content Specifications	Total number of items within the target range
	Number of operational items within the target range
	Number of field-test items within the target range
	Number of items per instructional area within the target range, evaluated separately for each instructional area
	Percentage of target-course items meets or exceeds the required minimum.
	Distribution of target-course and prerequisite items aligns with the student score distribution and reflects principled adaptation.
	Field-test items appear in designated positions.
Score Estimation	Theta ability parameter is recovered with $RMSE \leq 4$.
	Correlation between true and estimated scores is ≥ 0.95 .
Measurement Error	Median SEM is within the target range.
	SEM is similar across score deciles.
	Almost all students have SEM values at or below the maximum allowed SEM.
	Instructional area SEM is at or below the target SEM, evaluated for each instructional area.
Adaptivity	Item difficulty is appropriately aligned with momentary ability during the test event.

Simulation studies are conducted prior to publishing a test. After test administration, data from actual examinees are used to evaluate the extent to which the CAT satisfied blueprint, content, and statistical requirements. The results are summarized in a post-administration report.

6.3. Test Blueprints Have Been Satisfied

The MAP Growth adaptive test engine uses a heuristic approach to item selection. Test blueprints establish the target item counts for each content feature. However, item selection treats these targets as preferences instead of hard constraints. The engine prioritizes meeting blueprint targets, but it does not require every target to be met exactly. This approach helps avoid item starvation, which can occur when strict requirements are imposed and no items exist to meet them. By using preferences, each target has an acceptable range of values for an operational test event.

MAP Growth tests are tailored to content standards or frameworks by establishing customized blueprints. The number of items per instructional area is one aspect of the blueprints that may vary across tests. To evaluate the extent to which each test event satisfied its blueprint, a chi-

square goodness-of-fit test and corresponding effect size were calculated for each examinee. The statistic is shown in Equation 5.

$$\chi^2 = \sum_{k=1}^K \frac{(O_k - E_k)^2}{E_k} \quad (5)$$

where O_k is the observed count, and E_k is the expected count for category k . The statistic has $K - 1$ degrees of freedom. The chi-square statistic is a useful way to quantify the discrepancy between observed and expected counts under the null hypothesis that there is no difference. However, it cannot be used to prove the null hypothesis to be true, which is the goal for evaluating how well a test event satisfies blueprints (i.e., is no different from the blueprints). The chi-square test is converted to a Cramér's V effect size using Equation 6.

$$V = \sqrt{\frac{\chi^2}{N(K - 1)}} \quad (6)$$

where N is the total number of counts, and K is the total number of categories. The statistic ranges from 0 (no difference) to 1 (very substantial difference).

Various guidelines are available for interpreting the magnitude of Cramér's V . The ranges displayed in Table 6.2 were selected with the following rationale: In a four-category goodness-of-fit test with a total of 40 observations (e.g., Biology/Life Science tests), the practical meaning of Cramér's V can be expressed directly in terms of how far observed counts tend to deviate from their expected values. When expected counts are equal across categories (10, 10, 10, 10), a Cramér's V between 0.10 and 0.19 represents a small effect. If deviations are spread across categories, observed counts are typically about 2 to 3 counts away from 10 in each category, while if the discrepancy is concentrated, one category is usually about 3 to 6 counts away from 10 and the remaining categories differ by roughly 1 to 2 counts each. When Cramér's V increases to the 0.20–0.39 range, spread-out deviations are commonly on the order of about 3.5 to nearly 7 counts per category, whereas a concentrated pattern often shows one category differing by roughly 6 to 12 counts, with the others adjusting by about 2 to 4 counts each.

When expected counts are unequal (e.g., Geometry tests), the same effect-size ranges correspond to different raw count differences because smaller expected categories contribute more strongly to the chi-square statistic. For Cramér's V between 0.10 and 0.19, deviations are still small: if differences are spread out, the smallest category (expected 6) is typically about 1 to 2 counts away from expectation, while the larger categories (expected 10–13) are usually about 2 to 4 counts away. If the deviation is concentrated in one category, the main category is generally off by about 2.5 to 6 counts, with the other categories adjusting by roughly 1 to 3 counts each. For Cramér's V between 0.20 and 0.39, deviations become more clearly visible; when spread across categories, the smallest category is often about 3 to 5 counts off and the larger categories about 4 to 8 counts off, while in a concentrated pattern, one category may differ by roughly 5 to 15 counts, with the remaining categories typically shifting by about 2 to 5 counts each to maintain the total.

NWEA psychometricians compared the observed number of items per instructional area with the expected number documented in test blueprints by computing Cramér's V for each examinee. The percentage of test events in each effect size bin was then tabulated. As shown in Table 6.2, the substantial majority of MAP Growth course-specific test events in 2024–2025

had negligible effect sizes, indicating close alignment between observed and expected instructional area counts.

Across courses and terms, the percentages of test events with negligible effect sizes ranged from 87.33% for Biology/Life Science in spring to 97.60% for Integrated Math 3 in winter. Although the Biology/Life Science spring value was lower than the other course-term combinations, nearly all test events still fell in the negligible or small ranges, with only 0.05% in the medium range. This pattern suggests that spring Biology/Life Science showed somewhat more variation in blueprint alignment than other courses and terms, but the differences were generally minor and do not indicate substantial misalignment. Small effect sizes accounted for most of the remaining test events, ranging from 2.40% for Integrated Math 3 in winter to 12.61% for Biology/Life Science in spring. Medium effect sizes were rare, never exceeding 0.05% for any course or term, and no large or very large effect sizes were observed.

Results were generally stable across fall, winter, and spring administrations. Although Biology/Life Science showed a lower percentage of negligible effect sizes in spring than in fall or winter, nearly all spring test events were still classified as negligible or small. Taken together, these results indicate that MAP Growth course-specific test events closely met blueprint specifications across courses and testing terms.

Table 6.2. Percentage Distribution of Blueprint Alignment Effect Sizes by Course and Term

Course	Term	Negligible (< 0.10)	Small ($0.10-0.19$)	Medium ($0.20-0.39$)	Large ($0.40-0.59$)	Very large (≥ 0.60)
Algebra 1	Fall	93.33	6.65	0.02	0	0
	Winter	94.09	5.90	0.01	0	0
	Spring	93.76	6.22	0.02	0	0
Algebra 2	Fall	95.72	4.28	0	0	0
	Winter	97.15	2.85	0.01	0	0
	Spring	95.25	4.74	0.01	0	0
Geometry	Fall	94.69	5.30	0.01	0	0
	Winter	96.69	3.31	0	0	0
	Spring	93.27	6.71	0.02	0	0
Integrated Math 1	Fall	92.93	7.06	0.01	0	0
	Winter	91.03	8.95	0.02	0	0
	Spring	94.45	5.53	0.02	0	0
Integrated Math 2	Fall	95.63	4.36	0.01	0	0
	Winter	97.26	2.74	0	0	0
	Spring	95.80	4.20	0	0	0
Integrated Math 3	Fall	94.99	5.01	0	0	0
	Winter	97.60	2.40	0	0	0
	Spring	96.32	3.68	0	0	0
Biology/Life Science	Fall	97.01	2.98	0.01	0	0
	Winter	95.91	4.07	0.02	0	0
	Spring	87.33	12.61	0.05	0	0

6.4. The Domain Is Sampled for Efficient Tests and Aggregate Statistics

The MAP Growth item selection algorithm randomly samples items from each instructional area according to student ability. The content standards or frameworks for each instructional area are also randomly sampled because of the mapping of items to standards and the mapping of standards to instructional areas. It is a simple form of matrix sampling of test content (for a more complex form, see Zwick & Mislevy, 2011) that allows for shorter tests, especially in cases where the number of standards exceeds the number of items on the test. Individual students do not need to see items for every content standard; they need only see enough items to precisely estimate their location on the vertical scale. At higher levels of aggregation, where inferences are focused more on program and curriculum evaluation, content standards are more completely represented. That is, aggregate statistics (such as a district mean) will encompass the depth and breadth of the standards. Therefore, student scores and aggregate statistics (see Score Reports Facilitate the Interpretation and Use of Scores) represent the domain in a way that is efficient for testing and suitable for the primary inference about the information.

7. Student Scores and Item Difficulty Are on a Stable Vertical Scale

A vertical scale is a score scale that spans multiple grades and testing terms, allowing scores from one time point to be compared with scores from another. Vertical scaling is necessary for measuring student growth or change in achievement over time. MAP Growth uses subject-specific RIT scales across its assessment system, including Math, Reading, Language Usage, and Science. Course-specific math and science tests are reported on the same RIT scales as the corresponding grade-banded MAP Growth assessments, allowing course-specific scores and item difficulties within the same subject to be placed on a common scale across administrations.

Scaling is accomplished using the Rasch model (Rasch, 1980). MAP Growth person ability and item difficulty parameters are on the RIT scale (i.e., **Rasch unit** scale), which has a mean of 200 and a standard deviation of 10. Person ability scores (i.e., RIT scores) and item difficulty values on the RIT scale typically range from 100 to 350.

7.1. Operational Item Statistics

7.1.1. Classical Item Difficulty and Item Discrimination

Classical item statistics, including p values and point-biserial correlations, are used to evaluate the quality and functioning of individual test items. The p value represents the proportion of examinees who answered an item correctly and serves as an indicator of item difficulty, with lower values reflecting more-difficult items and higher values reflecting easier items. Items with moderate p values are typically preferred because they provide the most information across a wide range of student ability. The point-biserial correlation describes the relationship between performance on an individual item and performance on the test as a whole and is commonly interpreted as an index of item discrimination. Higher positive point-biserial values indicate that an item effectively differentiates between lower- and higher-performing students, while values near zero or negative may signal items that are poorly aligned with the construct being measured. Together, these statistics provide complementary evidence about whether items are appropriately targeted and functioning as intended within the assessment.

Table 7.1 shows that across courses and terms, mean p values generally ranged from 0.42 to 0.50, indicating that the item pools were moderately difficult overall. For Algebra 1, mean p values increased from 0.45 in fall to 0.48 in spring, with most items concentrated between 0.3 and 0.7 across terms. Algebra 2 and Geometry showed similar patterns, with mean p values increasing from 0.46 to 0.50 for Algebra 2 and from 0.45 to 0.49 for Geometry. These results suggest that the operational items were generally well targeted to the tested populations, with relatively few items in the very easy or very difficult ranges.

The Integrated Math courses also showed moderate item difficulty, though the number of items was smaller than Algebra 1, Algebra 2, and Geometry. Integrated Math 1 had mean p values ranging from 0.42 to 0.45 across terms, while Integrated Math 2 ranged from 0.45 to 0.46. Integrated Math 3 results were available for fall and winter only, with mean p values of 0.44 for both terms. Biology/Life Science showed a similar pattern to the math courses, with mean p values increasing from 0.45 in fall to 0.49 in spring and most items falling in the 0.3–0.7 range.

Table 7.2 shows that point-biserial correlations were positive and generally stable across courses and terms, indicating that items tended to discriminate appropriately between lower- and higher-performing students. For Algebra 1, mean point-biserial correlations ranged from 0.30 to 0.31, with the vast majority of items falling in the 0.2–0.4 range and a smaller proportion

in the 0.4–0.6 range. Algebra 2 showed slightly higher discrimination by spring, increasing from 0.31 in fall and winter to 0.33 in spring. Geometry had mean point-biserial correlations of 0.30 to 0.31, again with most items concentrated in the 0.2–0.4 range. Negative point-biserial values were essentially absent across these courses, supporting the conclusion that the items functioned as intended.

The Integrated Math courses demonstrated similarly acceptable discrimination. Mean point-biserial correlations ranged from 0.31 to 0.32 for Integrated Math 1, remained at 0.31 across terms for Integrated Math 2, and were 0.31 for Integrated Math 3 in both fall and winter. Although Integrated Math 3 had a smaller item count, its point-biserial values remained positive. Biology/Life Science also showed stable discrimination, with mean point-biserial correlations increasing from 0.30 in fall to 0.32 in spring. Across courses, the distributions were concentrated primarily in the 0.2–0.4 range, with additional items in the 0.4–0.6 range and very few items near zero or below zero. Overall, the course-specific item pools showed moderate difficulty and consistently positive item discrimination, providing evidence that the operational items were appropriately targeted and functioning effectively across administrations.

Table 7.1. Summary of Item *P* Values

Course	Term	N	Mean	SD	P-Value Range									
					[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
Algebra 1	Fall	1,761	0.45	0.11	0	0.28	6.25	25.04	36.23	23.34	7.89	0.85	0.11	0
	Winter	1,776	0.47	0.12	0.11	0.68	6.19	22.52	32.26	23.99	11.20	2.87	0.17	0
	Spring	1,882	0.48	0.12	0	0.80	5.37	19.29	32.25	25.56	13.12	3.45	0.16	0
Algebra 2	Fall	1,205	0.46	0.11	0.08	0.08	4.98	24.15	38.42	21.91	8.13	1.91	0.33	0
	Winter	1,129	0.48	0.12	0.09	0.44	5.23	21.26	31.98	24.98	9.83	5.58	0.62	0
	Spring	1,256	0.50	0.13	0.08	0.32	4.54	18.23	29.78	25.88	14.97	5.33	0.88	0
Geometry	Fall	1,302	0.45	0.11	0	0.84	6.45	23.12	39.40	21.43	5.99	1.84	0.92	0
	Winter	1,391	0.47	0.12	0	0.58	4.82	23.51	33.64	23.87	10.50	2.30	0.79	0
	Spring	1,577	0.49	0.12	0	1.14	4.82	17.06	31.20	27.96	14.27	3.23	0.32	0
Integrated Math 1	Fall	268	0.42	0.10	0	0.75	10.07	32.46	37.69	15.67	2.99	0.37	0	0
	Winter	245	0.44	0.10	0	0.82	6.12	27.35	40.00	20.00	4.49	1.22	0	0
	Spring	360	0.45	0.11	0	1.39	6.67	24.44	36.94	20.28	8.89	1.39	0	0
Integrated Math 2	Fall	278	0.45	0.11	0	0.72	6.47	28.06	31.65	21.58	8.99	2.16	0.36	0
	Winter	210	0.46	0.11	0	0.48	6.67	22.86	33.81	22.86	10.95	2.38	0	0
	Spring	302	0.46	0.12	0	0.33	6.29	25.17	32.45	21.85	10.93	2.65	0.33	0
Integrated Math 3	Fall	153	0.44	0.11	0	1.31	8.50	28.76	32.68	22.88	4.58	0	1.31	0
	Winter	142	0.44	0.11	0	0.70	9.86	29.58	30.99	21.83	5.63	0.70	0.70	0
Biology/ Life Science	Fall	652	0.45	0.09	0	0.46	4.91	22.70	43.10	23.01	5.37	0.46	0	0
	Winter	681	0.48	0.10	0	0.73	2.35	18.50	39.79	27.31	9.25	2.06	0	0
	Spring	663	0.49	0.12	0	0.75	2.87	17.80	33.63	26.85	13.88	3.32	0.90	0

Note. SD = standard deviation; Item-level *p*-value summary includes items meeting the minimum response criterion used for item-level analyses: at least 1,000 usable responses within each course and term, except for Integrated Math 1, Integrated Math 2, and Integrated Math 3, for which a minimum of 300 usable responses was used due to lower item-level response volume.

Table 7.2. Summary of Item Point-Biserial Correlations

Course	Term	N	Mean	SD	Point-Biserial Range									
					[-1.0, -0.8]	(-0.8, -0.6]	(-0.6, -0.4]	(-0.4, -0.2]	(-0.2, 0.0]	(0.0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1.0]
Algebra 1	Fall	1,761	0.30	0.06	0	0	0	0	0	5.28	89.04	5.68	0	0
	Winter	1,776	0.30	0.06	0	0	0	0	0	5.01	89.36	5.57	0.06	0
	Spring	1,882	0.31	0.07	0	0	0	0	0	3.67	88.47	7.86	0	0
Algebra 2	Fall	1,205	0.31	0.06	0	0	0	0	0.08	4.15	87.72	8.05	0	0
	Winter	1,129	0.31	0.07	0	0	0	0	0	3.54	87.78	8.68	0	0
	Spring	1,256	0.33	0.07	0	0	0	0	0.08	2.31	86.23	11.39	0	0
Geometry	Fall	1,302	0.30	0.07	0	0	0	0	0.08	8.68	83.95	7.30	0	0
	Winter	1,391	0.31	0.07	0	0	0	0	0	6.33	84.97	8.70	0	0
	Spring	1,577	0.31	0.07	0	0	0	0	0	5.07	86.18	8.75	0	0
Integrated Math 1	Fall	268	0.31	0.07	0	0	0	0	0	5.97	83.96	10.07	0	0
	Winter	245	0.31	0.07	0	0	0	0	0	6.12	80.41	13.47	0	0
	Spring	360	0.32	0.07	0	0	0	0	0.28	4.17	83.89	11.67	0	0
Integrated Math 2	Fall	278	0.31	0.08	0	0	0	0	0	7.19	82.01	10.79	0	0
	Winter	210	0.31	0.07	0	0	0	0	0	5.71	82.38	11.90	0	0
	Spring	302	0.31	0.08	0	0	0	0	0	8.28	82.78	8.94	0	0
Integrated Math 3	Fall	153	0.31	0.08	0	0	0	0	0	10.46	75.82	13.73	0	0
	Winter	142	0.31	0.09	0	0	0	0	0	12.68	69.72	17.61	0	0
Biology/ Life Science	Fall	652	0.30	0.06	0	0	0	0	0	4.60	89.11	6.29	0	0
	Winter	681	0.31	0.06	0	0	0	0	0	2.94	91.48	5.58	0	0
	Spring	663	0.32	0.07	0	0	0	0	0.15	3.47	88.24	8.14	0	0

Note. SD = standard deviation; Item-level point-biserial summary includes items meeting the minimum response criterion used for item-level analyses: at least 1,000 usable responses within each course and term, except for Integrated Math 1, Integrated Math 2, and Integrated Math 3, for which a minimum of 300 usable responses was used due to lower item-level response volume.

7.2. Item Fit

Item OUTFIT is a fit statistic used to evaluate how well observed item responses align with the expectations of the measurement model. The expected value of the OUTFIT is 1.0, with values close to 1.0 indicating good fit. Values between 0.5 and 1.5 are generally considered acceptable, while values outside this range indicate potential item misfit.

One caveat about item fit statistics in a computer adaptive test is that items are selected according to the momentary person ability value. The momentary value for items at the beginning of the test can be very different from the final person ability value computed at the end of the test. As such, more misfit is expected when using the final ability estimate in calculations than when using the momentary ability estimate. OUTFIT values are computed using the final ability estimate.

7.2.1. Cross-Sectional Item Fit

Item fit statistics are computed for each item by term and subject. Figure 7.1 shows the distribution of these values across the course-specific assessments. Across courses and terms, the distributions were centered near 1.0, indicating that most items fit the measurement model as expected. The interquartile ranges were relatively narrow for most courses, suggesting limited variability in item fit values. Biology/Life Science showed the most compact distributions, while Integrated Math 3 showed somewhat greater variability, particularly in spring. Across courses, spring administrations tended to show slightly wider distributions than fall or winter.

A limitation of the term-by-term results in Figure 7.1 is that they do not indicate whether the same items showed misfit across multiple terms. Items that show misfit in one term may fit adequately in another term, or the same items may show repeated misfit across administrations. For this reason, longitudinal fit analyses were conducted to evaluate whether items maintained fit across fall, winter, and spring administrations.

7.2.2. Longitudinal Item Fit

Longitudinal item fit statistics are computed for operational items administered in fall, winter, and spring with 250 or more responses in each term. To be included, an item had to be administered operationally (rather than as a field-test item) in all three terms. Fit statistics are expected to remain within the acceptable range at each term; items with fit values consistently outside the acceptable range may warrant additional review or recalibration.

Table 7.3 summarizes the number and percentage of items in each course with no evidence of misfit or with misfit in one, two, or three terms. Across course-specific assessments, most items showed no misfit in any term. The percentage of items with no misfit ranged from 62.07% for Integrated Math 3 to 95.23% for Biology/Life Science. Algebra 2, Geometry, Integrated Math 1, and Integrated Math 2 each had more than 73% of items with no misfit, while Algebra 1 had 67.78%.

Items showing misfit in only one term accounted for a smaller proportion of items, ranging from 4.13% for Biology/Life Science to 21.98% for Integrated Math 3. Repeated misfit across two or three terms was less common. The percentage of items with misfit in two terms ranged from 0.28% for Biology/Life Science to 9.91% for Integrated Math 3, and the percentage with misfit in all three terms ranged from 0.37% for Biology/Life Science to 6.03% for Integrated Math 3. Overall, these results indicate that most course-specific items fit well across terms, with relatively few items showing persistent misfit.

Figure 7.1. Item OUTFIT Statistics by Course and Term

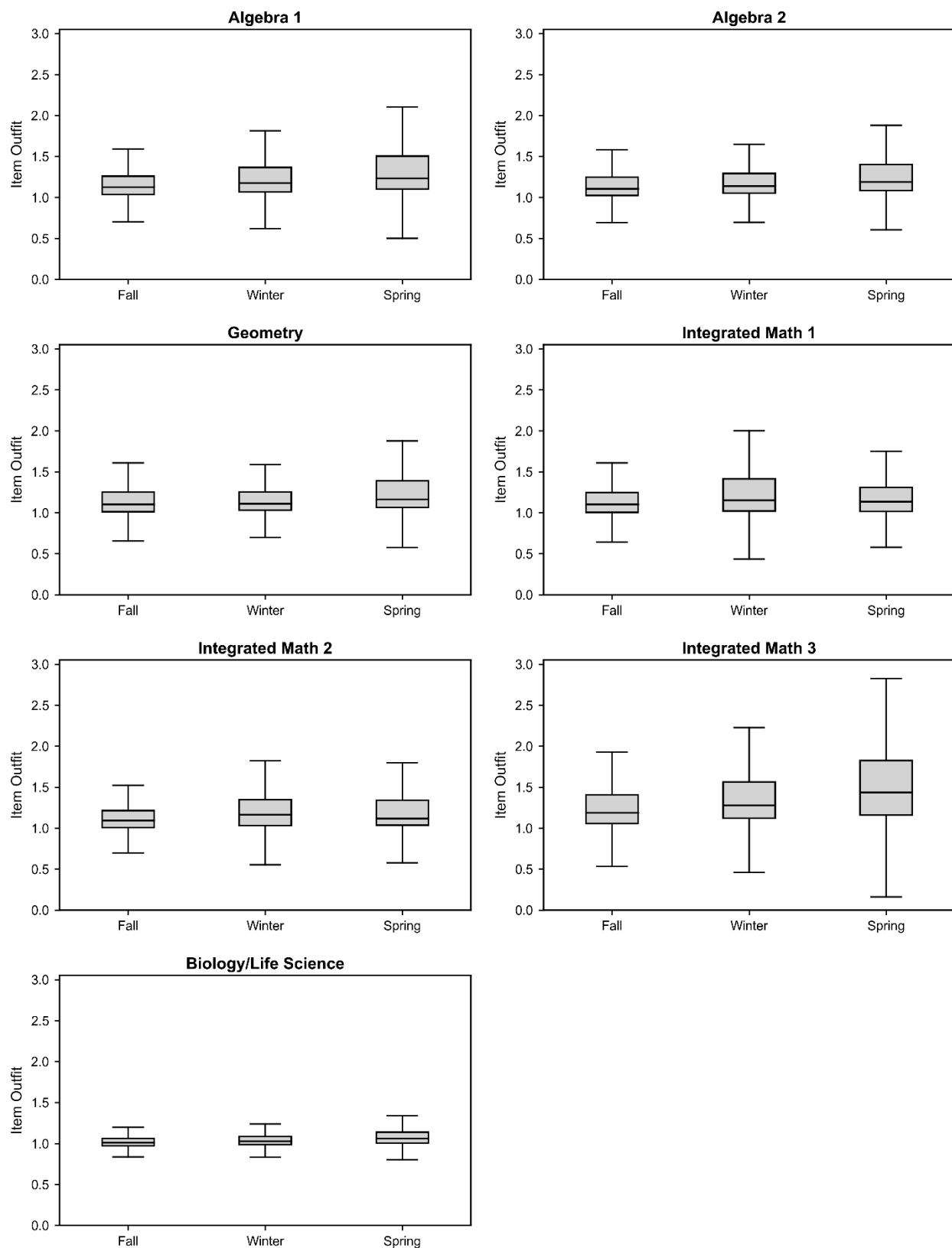


Table 7.3. Item Misfit Summary by Course

Course	Number of Terms with Item Misfit	Number of Items ^a	Percentage of Items
Algebra 1	0	2,711	67.78
	1	790	19.75
	2	318	7.95
	3	181	4.53
Algebra 2	0	1,713	75.60
	1	354	15.62
	2	146	6.44
	3	53	2.34
Geometry	0	1,699	73.42
	1	380	16.42
	2	180	7.78
	3	55	2.38
Integrated Math 1	0	318	77.37
	1	55	13.38
	2	23	5.60
	3	15	3.65
Integrated Math 2	0	306	81.82
	1	39	10.43
	2	20	5.35
	3	9	2.41
Integrated Math 3	0	144	62.07
	1	51	21.98
	2	23	9.91
	3	14	6.03
Biology/ Life Science	0	1,038	95.23
	1	45	4.13
	2	3	0.28
	3	4	0.37

^a Items are counted multiple times when observed in more than one term.

For MAP Growth more generally, item parameter drift studies are conducted on a regular basis for operational items. Items flagged as drifting are recalibrated to the scale, and their operational item difficulty values are updated in the item pool if the item passes review criteria after recalibration. For example, He (2022) analyzed over 9,000 operational items and found that between 4.2% and 6.4% of the items across subjects demonstrated drift and needed to have their parameters updated. Overall, the item and test-level analysis in He’s study showed that the scales remained stable.

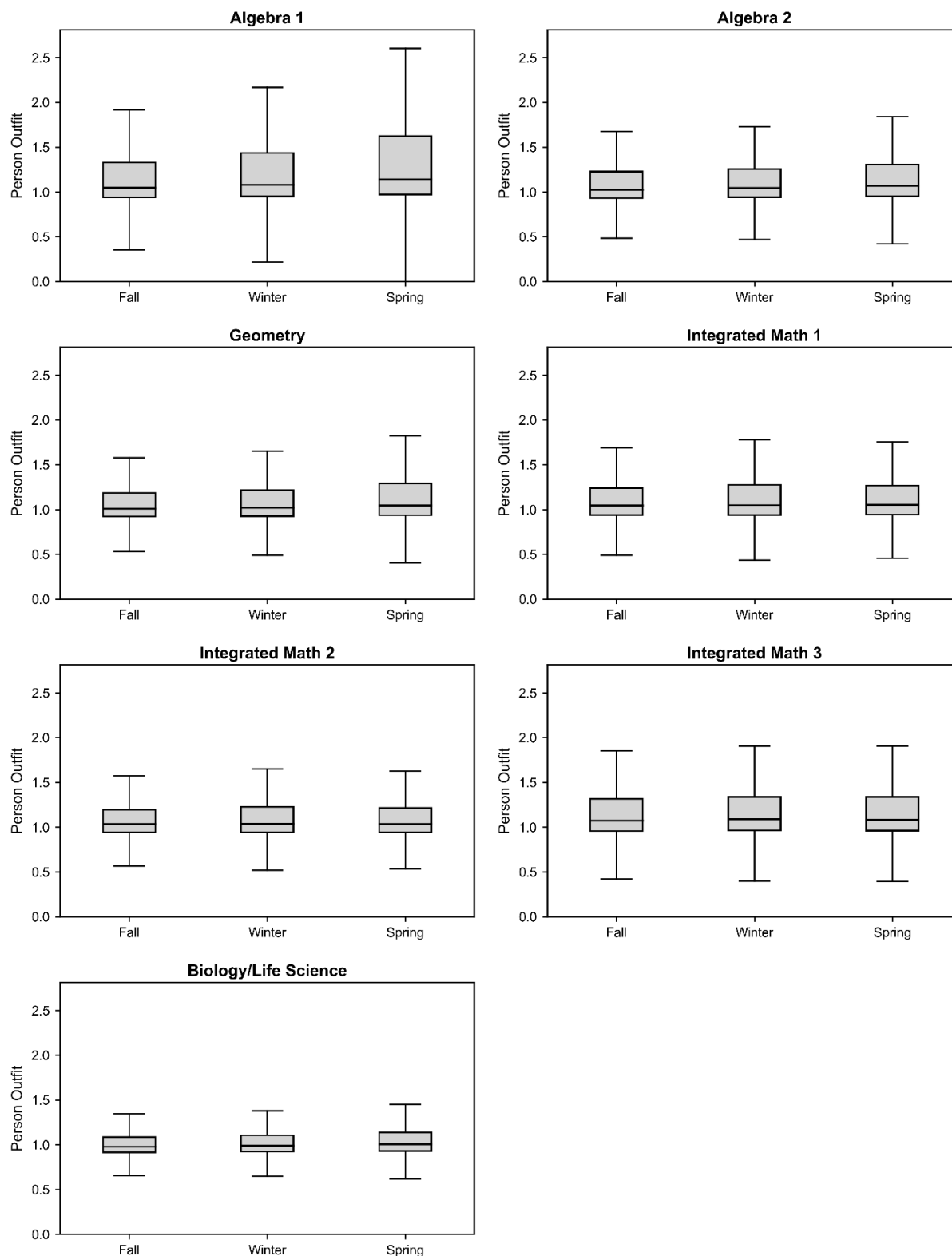
7.3. Examinee Fit

A Rasch OUTFIT statistic computed for an examinee indicates the degree to which the examinee’s response pattern fits the measurement model, given the items administered. Person OUTFIT is interpreted in the same manner as item OUTFIT: the expected value is 1.0, with values between 0.5 and 1.5 generally considered acceptable.

As shown in Figure 7.2, person OUTFIT values were generally centered near 1.0 across course-specific assessments and testing terms. The median values were close to 1.0 for each course, and the middle 50% of values typically fell within a narrow range around 1.0. Biology/Life Science showed the most compact distributions, while Algebra 1 showed somewhat greater

variability, particularly in spring. Overall, the distributions indicate that examinee response patterns generally fit the measurement model well across courses and administrations.

Figure 7.2. Examinee OUTFIT Statistics by Course and Term



7.4. Examinee Scores and Item Difficulty Increase by Grade and Show Variability

Course-specific assessments are designed to measure achievement within defined course contexts rather than across successive grade levels. In the Rasch model, examinee scores and item difficulties are placed on the same RIT scale, which makes it possible to evaluate how well the difficulty of administered items aligns with the performance levels of students taking each assessment. When items are well matched to examinees, average item difficulty should be close to the average student score, and the variability of item difficulties should be similar to the variability of student scores.

Table 7.4 compares student RIT scores and item difficulties for course-specific assessments. Across courses, average item difficulties were closely aligned with average student scores, with item means slightly higher than student means in each course. Differences between student means and item means were small, ranging from approximately 1 to 3 RIT points. The standard deviations of student scores and item difficulties were also similar within each course, indicating that the range of item difficulties was generally well matched to the range of student achievement. These results provide evidence that the course-specific item pools are appropriately targeted to the students taking each assessment.

Table 7.4. Comparison of Examinee RIT Scores to Item Difficulties

Course	Student		Item	
	Mean	SD	Mean	SD
Algebra 1	227.78	20.90	230.10	19.95
Algebra 2	242.87	20.69	244.18	18.66
Geometry	233.88	19.83	235.50	18.82
Integrated Math 1	225.82	19.77	228.26	18.33
Integrated Math 2	234.42	18.90	236.52	17.57
Integrated Math 3	242.22	20.80	244.29	19.08
Biology/Life Science	216.63	15.72	217.84	14.46

Note. SD = standard deviation

7.5. Conditional Standard Error of Measurement

In item response theory, the amount of information an item provides about a student's ability (i.e., θ) is quantified by the *item information function*. The larger the value, the more informative the item. For the Rasch model, this function is $I_{ij}(\theta) = P_{ij}(\theta)[1 - P_{ij}(\theta)]$. The *test information function* is the sum of item information values for all J items on a student's test, as shown in Equation 7.

$$I_i(\theta) = \sum_{j=1}^J I_{ij}(\theta) \quad (7)$$

Test information indicates the precision of a test score. The larger the test information function, the more precisely a student's score is estimated. The test information function is like a person-specific reliability such that the larger the value, the more reliable the estimate.

Measurement error is the amount of random error associated with estimating a student's score. The standard error of measurement (SEM) is conditional on student ability (i.e., θ) and describes the amount of random error associated with estimating a student's score. The conditional standard error of measurement (CSEM) is shown in Equation 8.

$$CSEM_i(\theta) = \frac{1}{\sqrt{I_i(\theta)}} \quad (8)$$

The CSEM is inversely related to the test information function. The larger the test information, the lower the CSEM. Because test information increases as the number of items on the test increases, the CSEM becomes smaller as the number of items on the test increases.

7.5.1. CSEM Results

A MAP Growth course-specific test event is designed to have a target CSEM of approximately 3.3 RIT points. For a given test, the target CSEM is the same for all students, and the observed CSEMs are expected to be relatively stable across the score scale. Figure 7.3 **Error!**

Reference source not found. displays CSEM distributions for students taking course-specific assessments, grouped by RIT score decile. Across courses, CSEM values were generally stable across deciles, with the smallest values typically observed near the middle of the score distribution and somewhat larger values observed at the lower and upper extremes.

Table 7.5 provides numerical summaries for the 1st, 5th, and 10th score deciles by course and term. Across courses and administrations, average CSEM values were close to the target value of 3.3 RIT points. The 5th decile values were highly consistent, ranging from 3.31 to 3.40 across courses and terms. CSEM values were higher in the 1st decile, generally ranging from 3.64 to 4.08, and were also somewhat higher in the 10th decile, ranging from 3.36 to 3.76. Biology/Life Science tended to have slightly lower CSEM values in the 1st decile than the math courses, while spring Algebra 2 showed the highest 10th-decile value. Overall, the results indicate that course-specific assessments maintained relatively consistent measurement precision across the score scale, with modestly higher measurement error at the extremes.

As shown in Table 7.6, CSEM values were generally similar across course, demographic group, and term, with the clearest differences observed by score decile. CSEM values were slightly higher in the lower and upper deciles, indicating somewhat greater measurement error at the extremes of the score distribution. Differences across other factors were small and, when present, were typically only a fraction of a RIT point.

Figure 7.3. CSEM Distribution by Course and Score Decile

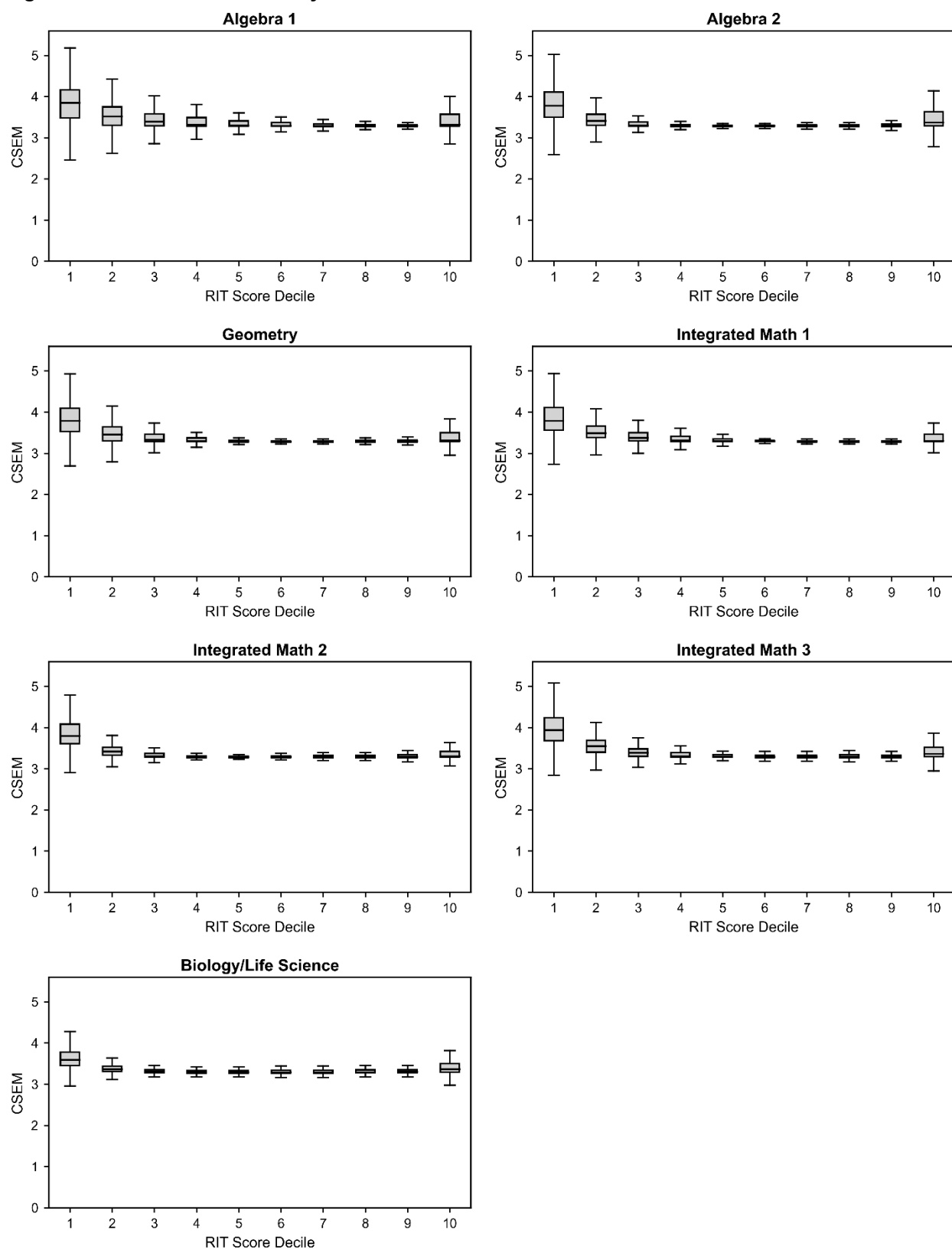


Table 7.5. CSEM Summary by Course, Term, and Score Decile

Course	Fall Score Deciles			Winter Score Deciles			Spring Score Deciles		
	1st	5th	10th	1st	5th	10th	1st	5th	10th
Algebra 1	3.89	3.34	3.37	3.88	3.38	3.49	3.90	3.40	3.61
Algebra 2	3.88	3.31	3.42	3.85	3.32	3.51	3.87	3.33	3.76
Geometry	3.80	3.31	3.36	3.91	3.32	3.44	3.97	3.32	3.53
Integrated Math 1	3.91	3.35	3.39	3.88	3.34	3.48	3.91	3.33	3.47
Integrated Math 2	3.93	3.31	3.39	3.86	3.33	3.49	3.82	3.31	3.39
Integrated Math 3	3.94	3.31	3.44	3.98	3.38	3.45	4.08	3.32	3.52
Biology/Life Science	3.66	3.33	3.36	3.69	3.33	3.43	3.64	3.29	3.59

Table 7.6. CSEM Summary by Course, Demographic Group, Term, and Score Decile

Course	Demographic Group	Fall Score Deciles			Winter Score Deciles			Spring Score Deciles		
		1st	5th	10th	1st	5th	10th	1st	5th	10th
Algebra 1	Female	3.84	3.33	3.38	3.83	3.37	3.50	3.86	3.39	3.62
	Male	3.89	3.34	3.37	3.89	3.36	3.48	3.91	3.39	3.60
	White	3.86	3.33	3.34	3.87	3.37	3.42	3.87	3.38	3.52
	Black	3.81	3.34	3.46	3.80	3.36	3.61	3.84	3.39	3.72
	Asian	3.82	3.34	3.34	3.76	3.35	3.38	3.80	3.37	3.50
	Hispanic	3.91	3.34	3.42	3.90	3.37	3.57	3.91	3.40	3.68
	AI/AN	3.87	3.33	3.40	3.85	3.35	3.52	3.85	3.39	3.62
	NH/PI	3.90	3.35	3.40	3.79	3.38	3.44	3.80	3.38	3.53
	Multiple	3.91	3.34	3.37	3.93	3.36	3.52	4.01	3.40	3.65
	Other	3.80	3.35	3.37	3.78	3.38	3.46	3.88	3.44	3.68
Algebra 2	Female	3.93	3.31	3.42	3.92	3.31	3.53	3.92	3.32	3.74
	Male	3.98	3.31	3.42	3.98	3.32	3.50	3.97	3.33	3.77
	White	3.94	3.31	3.37	3.92	3.31	3.43	3.89	3.31	3.65
	Black	3.93	3.31	3.53	3.94	3.32	3.70	3.96	3.34	3.84
	Asian	3.88	3.32	3.47	3.85	3.32	3.47	3.87	3.33	3.79
	Hispanic	3.98	3.31	3.45	3.98	3.32	3.57	3.97	3.33	3.78
	AI/AN	3.96	3.31	3.51	3.92	3.31	3.68	3.84	3.32	3.81
	NH/PI	3.89	3.30	3.38	3.79	3.32	3.51	3.67	3.35	3.88
	Multiple	3.99	3.32	3.42	3.96	3.31	3.53	4.05	3.33	3.76
	Other	3.87	3.33	3.37	3.80	3.34	3.55	3.80	3.37	3.95
Geometry	Female	3.78	3.31	3.37	3.89	3.32	3.46	3.94	3.32	3.55
	Male	3.83	3.31	3.35	3.93	3.32	3.42	3.99	3.32	3.51
	White	3.78	3.31	3.33	3.88	3.31	3.39	3.92	3.31	3.47
	Black	3.78	3.31	3.47	3.88	3.32	3.61	3.94	3.33	3.69
	Asian	3.86	3.31	3.35	3.93	3.31	3.36	3.99	3.32	3.43
	Hispanic	3.83	3.31	3.40	3.95	3.32	3.51	4.01	3.32	3.59
	AI/AN	3.82	3.32	3.36	3.94	3.31	3.47	3.98	3.31	3.53
	NH/PI	3.75	3.30	3.37	3.93	3.32	3.44	4.01	3.28	3.50
	Multiple	3.75	3.31	3.37	3.82	3.32	3.48	4.01	3.32	3.58

Course	Demographic Group	Fall Score Deciles			Winter Score Deciles			Spring Score Deciles		
		1st	5th	10th	1st	5th	10th	1st	5th	10th
	Other	3.78	3.32	3.33	3.84	3.32	3.39	3.89	3.31	3.48
Integrated Math 1	Female	3.91	3.35	3.41	3.89	3.33	3.51	3.92	3.33	3.49
	Male	3.91	3.35	3.38	3.86	3.34	3.46	3.90	3.32	3.45
	White	3.90	3.33	3.36	3.85	3.33	3.38	3.91	3.32	3.37
	Black	3.77	3.34	3.48	3.84	3.33	3.64	3.73	3.35	3.68
	Asian	4.02	3.36	3.35	4.13	3.36	3.36	4.02	3.34	3.39
	Hispanic	3.94	3.36	3.40	3.91	3.33	3.51	3.96	3.33	3.51
	AI/AN	4.98	3.46	3.61	3.76	3.41	3.72	3.77	3.30	3.70
	Multiple	3.82	3.36	3.45	3.75	3.30	3.59	3.86	3.31	3.45
	Other	3.83	3.32	3.37	3.78	3.39	3.47	3.75	3.31	3.51
Integrated Math 2	Female	3.93	3.32	3.39	3.86	3.33	3.52	3.82	3.31	3.42
	Male	3.93	3.31	3.39	3.86	3.34	3.46	3.82	3.32	3.37
	White	3.98	3.30	3.37	3.85	3.35	3.39	3.81	3.30	3.34
	Black	3.92	3.36	3.45	3.76	3.31	3.58	3.66	3.30	3.66
	Asian	4.01	3.32	3.35	3.81	3.33	3.41	3.97	3.33	3.35
	Hispanic	3.91	3.31	3.38	3.88	3.34	3.61	3.86	3.32	3.42
	AI/AN	3.66	3.30	3.30	3.40	3.35	3.52	3.68	3.28	3.40
	Multiple	3.91	3.38	3.49	3.73	3.31	3.42	3.79	3.31	3.36
	Other	3.81	3.37	3.64	3.88	3.32	3.51	3.83	3.29	3.43
Integrated Math 3	Female	3.97	3.31	3.47	3.98	3.38	3.50	4.02	3.33	3.54
	Male	3.92	3.32	3.43	3.99	3.38	3.41	4.14	3.31	3.50
	White	3.79	3.31	3.42	3.90	3.35	3.40	3.96	3.30	3.49
	Black	4.10	3.31	3.50	3.88	3.36	3.52	4.10	3.35	3.31
	Asian	4.11	3.29	3.37	4.21	3.45	3.37	4.23	3.29	3.44
	Hispanic	4.02	3.32	3.49	4.02	3.38	3.55	4.11	3.32	3.54
	AI/AN	3.78	3.29	3.66	3.84	3.35	3.52	3.59	3.29	3.58
	Multiple	3.94	3.29	3.55	4.36	3.38	3.43	4.01	3.30	3.47
	Other	3.84	3.31	3.51	3.92	3.41	3.54	4.11	3.34	3.63
Biology/Life Science	Female	3.64	3.33	3.36	3.67	3.32	3.43	3.62	3.29	3.57
	Male	3.67	3.33	3.36	3.71	3.33	3.43	3.65	3.30	3.61
	White	3.65	3.33	3.35	3.66	3.32	3.40	3.60	3.30	3.58
	Black	3.62	3.34	3.40	3.65	3.33	3.49	3.62	3.30	3.60
	Asian	3.66	3.33	3.36	3.74	3.32	3.39	3.65	3.30	3.69
	Hispanic	3.68	3.33	3.36	3.72	3.32	3.46	3.65	3.29	3.58
	AI/AN	3.67	3.32	3.37	3.74	3.33	3.48	3.69	3.29	3.58
	NH/PI	3.64	3.33	3.35	3.70	3.34	3.50	3.59	3.28	3.55
	Multiple	3.63	3.33	3.35	3.69	3.33	3.43	3.68	3.30	3.61
	Other	3.59	3.32	3.35	3.61	3.31	3.39	3.63	3.30	3.62

Note. AI/AN = American Indian or Alaska Native; NH/PI = Native Hawaiian or Pacific Islander

7.6. Score Reliability

The CSEM is dependent on the scale on which scores are reported. For example, if a score scale were multiplied by 10, the corresponding CSEM values would also be multiplied by 10. Score reliability overcomes this limitation because it is a scale-free description of the precision within a set of scores. Changing the scale does not affect the value of a reliability estimate, which allows reliability to be used to compare score precision across assessments.

Test information and CSEM values are computed for individual students, whereas a reliability coefficient is computed for a group of students to describe the consistency of scores for that group. That is, reliability refers to the consistency of test scores for a group of examinees. It is defined as the ratio of true score variance to observed score variance, $\rho_{XT}^2 = \sigma_T^2 / (\sigma_T^2 + \sigma_E^2)$. Reliability decreases as measurement error increases. Reliability estimates range from 0 to 1. Values closer to 1 indicate more consistent scores and less measurement error, whereas values closer to 0 indicate little true score variance, substantial error variance, or both.

There are multiple types of reliability estimates, and each reflects the influence of a particular source of error on score consistency. *Test-retest* reliability is a coefficient of stability that captures the influence of time on score consistency. It is calculated as the correlation between scores from the same students on two testing occasions. A variation of this estimate is test-retest with alternate forms, where the test forms given on each occasion are very similar but not identical. This type of reliability reflects two sources of error: time and item sampling.

Internal consistency is another type of reliability estimate. It reflects the consistency of scores in the presence of error due to item sampling. In item response theory (IRT), *marginal reliability* is commonly used to estimate internal consistency for a group of students. It is calculated as shown in Equation 9.

$$\hat{\rho}_{TX}^2 = \frac{\sigma_\theta^2}{\sigma_\theta^2 + E[SEM(\theta)]} \quad (9)$$

where σ_θ^2 is the variance of IRT-based scores for a group of examinees, and $E[SEM(\theta)]$ is the expected value of examinee-specific SEM values taken over the group of examinees.

In the definition of the reliability coefficient, the numerator is the variance of true scores for a group of examinees, σ_T^2 . As such, the reliability coefficient depends on the group of students for which the reliability estimate is calculated. A group with greater true score variance, such as all students taking Algebra 1 nationally, will generally have higher reliability estimates than a more homogenous group with less true score variance, such as students in the same district taking Algebra 1 after completing the same accelerated algebra curriculum. For this reason, it is necessary to estimate reliability for major subgroups of students taking a test, including demographic groups such as race and ethnicity. Reliability estimates should be high and of similar magnitude across groups.

7.6.1. Marginal Reliability Results

Marginal reliability estimates were examined by course, term, and demographic group for the MAP Growth course-specific assessments. Results indicate consistently strong score reliability across the studied factors, with estimates generally in the mid-to-high 0.90s.

At the course level, marginal reliability estimates were high across fall, winter, and spring, as shown in Table 7.7. Reliability ranged from 0.94 to 0.98 across courses and terms. Math course reliabilities were consistently strong, ranging from 0.96 to 0.98. Biology/Life Science reliabilities were slightly lower but still high, increasing from 0.94 in fall to 0.96 in spring. Across courses, reliability was generally stable across terms, with a slight tendency for winter and spring estimates to be equal to or higher than fall estimates.

Table 7.8 shows that marginal reliability estimates were also consistently high across demographic groups. For Algebra 1, Algebra 2, Geometry, Integrated Math 1, Integrated Math 2, and Integrated Math 3, subgroup reliability estimates were generally between 0.95 and 0.98 across terms. Biology/Life Science estimates were somewhat lower than the math courses, ranging from 0.93 to 0.96 across demographic groups and terms, but remained within a strong range. Differences across demographic groups were small, typically no more than one or two hundredths within a course and term.

Overall, these results indicate that MAP Growth course-specific assessments provide reliable scores across courses, testing terms, and reported demographic groups. The consistency of reliability estimates supports the use of these scores for interpreting student achievement in course-based contexts.

As shown in Appendix A: Marginal Reliability by Course, State, and Term, marginal reliability estimates by state and term were consistently high, generally falling in the upper 0.90s. Appendix B: Marginal Reliability by Course, Instructional Area, and Term provides additional marginal reliability results by course, instructional area, and term. Results in both appendices were omitted when the sample size was below 1,000 in any term.

Table 7.7. Marginal Reliability by Course and Term

Course	Fall		Winter		Spring	
	N	Reliability	N	Reliability	N	Reliability
Algebra 1	214,951	0.96	175,227	0.97	165,273	0.97
Algebra 2	133,474	0.97	101,600	0.97	106,177	0.98
Geometry	179,630	0.96	141,850	0.97	136,628	0.97
Integrated Math 1	13,056	0.96	8,092	0.97	10,452	0.97
Integrated Math 2	9,781	0.96	6,003	0.97	8,423	0.97
Integrated Math 3	4,891	0.97	3,894	0.97	3,215	0.98
Biology/Life Science	116,790	0.94	95,364	0.95	96,354	0.96

Table 7.8. Marginal Reliability by Course, Demographic Group, and Term

Course	Demographic Group	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Algebra 1	Female	102,452	0.96	83,717	0.97	79,175	0.97
	Male	110,674	0.96	90,066	0.97	84,710	0.97
	White	70,118	0.96	56,926	0.96	52,213	0.97
	Black	41,253	0.96	34,916	0.97	29,586	0.98
	Asian	8,966	0.97	7,716	0.97	7,314	0.98
	Hispanic	73,674	0.96	59,622	0.97	58,799	0.97
	AI/AN	2,971	0.96	2,521	0.97	2,135	0.97

Course	Demographic Group	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
	Multiple	9,758	0.96	7,696	0.97	8,665	0.97
	Other	5,886	0.96	4,009	0.96	4,827	0.97
Algebra 2	Female	66,432	0.97	50,758	0.97	52,986	0.98
	Male	66,972	0.97	50,788	0.97	53,152	0.98
	White	43,110	0.96	33,206	0.97	32,106	0.97
	Black	17,658	0.97	14,798	0.97	13,987	0.98
	Asian	11,876	0.97	9,250	0.97	9,271	0.97
	Hispanic	48,983	0.96	35,675	0.97	40,202	0.97
	AI/AN	1,831	0.96	1,524	0.97	1,209	0.98
	Multiple	5,705	0.97	4,458	0.97	6,130	0.98
	Other	3,967	0.96	2,439	0.97	3,042	0.98
Geometry	Female	88,781	0.96	70,296	0.96	67,432	0.97
	Male	90,743	0.96	71,493	0.97	69,117	0.98
	White	66,016	0.96	52,065	0.96	48,482	0.97
	Black	26,849	0.96	22,101	0.97	18,972	0.98
	Asian	12,399	0.96	10,405	0.97	9,663	0.97
	Hispanic	59,300	0.96	46,385	0.96	47,130	0.97
	AI/AN	2,281	0.96	1,770	0.97	1,427	0.97
	Multiple	7,961	0.96	6,081	0.97	7,363	0.98
	Other	4,343	0.95	2,692	0.96	3,246	0.97
Integrated Math 1	Female	6,222	0.96	3,904	0.97	5,109	0.97
	Male	6,827	0.97	4,180	0.97	5,336	0.97
	White	3,424	0.96	1,639	0.97	2,459	0.97
	Hispanic	5,928	0.95	3,896	0.96	5,286	0.97
Integrated Math 2	Female	4,728	0.96	2,934	0.97	4,131	0.97
	Male	5,045	0.96	3,064	0.97	4,284	0.97
	White	2,992	0.96	1,261	0.97	2,225	0.97
	Hispanic	4,444	0.95	3,140	0.96	4,065	0.97
Integrated Math 3	Female	2,395	0.96	1,897	0.97	1,575	0.97
	Male	2,493	0.97	1,994	0.97	1,637	0.98
	Hispanic	1,749	0.96	1,454	0.97	1,400	0.98
Biology/ Life Science	Female	57,023	0.93	46,628	0.95	47,026	0.96
	Male	59,760	0.94	48,728	0.95	49,326	0.96
	White	36,408	0.94	30,039	0.95	29,853	0.96
	Black	17,282	0.93	14,705	0.95	13,872	0.96
	Asian	5,451	0.94	4,963	0.95	4,495	0.95
	Hispanic	45,938	0.93	36,695	0.95	38,754	0.96
	AI/AN	2,001	0.93	1,402	0.94	1,332	0.96
	Multiple	6,230	0.94	5,127	0.95	6,629	0.96
	Other	3,257	0.94	2,242	0.94	1,218	0.95

Note. AI/AN = American Indian or Alaska Native. Results for a demographic group were omitted if the sample size for any term was less than 1,000.

7.6.2. Test-Retest Reliability Results

MAP Growth course-specific assessments may be administered multiple times during the school year, making test-retest reliability an important source of evidence for score stability over time. Because MAP Growth is adaptive, test-retest reliability is best interpreted as test-retest reliability with alternate forms. Students do not receive the same set of items across administrations; instead, each test event is comparable in content and structure but differs in the specific items administered and their difficulty levels. As a result, these estimates reflect the influence of both time and adaptive item selection.

Test-retest reliability was estimated using Pearson correlations between RIT scores from students who tested in two terms during the school year: fall–winter, fall–spring, and winter–spring. As shown in Table 7.9, test-retest reliability estimates were moderate to strong across courses. Fall–winter correlations ranged from 0.72 for Algebra 1 to 0.78 for Integrated Math 3. Winter–spring correlations ranged from 0.68 for Algebra 1 to 0.75 for Integrated Math 2. Fall–spring correlations were lower, ranging from 0.61 for Algebra 1 to 0.69 for Integrated Math 2. This pattern is expected because the fall–spring interval spans more instructional time, during which true changes in achievement are more likely to occur.

Table 7.10 shows test-retest reliability by course and demographic group. Results were generally consistent across reported subgroups, with the same overall pattern of higher correlations for shorter intervals and lower correlations for the longer fall–spring interval. Some subgroup variation was observed, particularly for smaller groups, but no broad pattern indicated systematically weak score stability for any demographic group across courses. Results for demographic groups were omitted when the sample size for any term was less than 1,000. Overall, the test-retest results support the temporal stability of MAP Growth course-specific scores while also reflecting expected changes in student achievement over the school year. These findings provide evidence that course-specific scores can be used to monitor student progress across testing terms and to support longitudinal interpretations of achievement within course-based contexts.

Table 7.9. Test-Retest Reliability by Course

Course	N	Fall–Winter	Fall–Spring	Winter–Spring
Algebra 1	104,774	0.72	0.61	0.68
Algebra 2	68,282	0.77	0.64	0.72
Geometry	89,077	0.76	0.66	0.73
Integrated Math 1	4,005	0.77	0.67	0.74
Integrated Math 2	3,001	0.76	0.69	0.75
Integrated Math 3	1,762	0.78	0.65	0.71
Biology/Life Science	64,678	0.75	0.64	0.70

Table 7.10. Test-Retest Reliability by Course and Demographic Group

Course	Demographic Group	N	Fall–Winter	Fall–Spring	Winter–Spring
Algebra 1	Female	50,820	0.71	0.60	0.68
	Male	53,933	0.73	0.62	0.68
	Asian	4,586	0.80	0.68	0.73

Course	Demographic Group	N	Fall–Winter	Fall–Spring	Winter–Spring
	Black	18,029	0.67	0.54	0.62
	AI/AN	1,389	0.67	0.58	0.65
	Hispanic	36,190	0.69	0.58	0.65
	Other	1,906	0.71	0.57	0.66
	Multiple	4,424	0.71	0.64	0.71
	White	34,755	0.76	0.67	0.73
Algebra 2	Female	34,177	0.74	0.62	0.71
	Male	34,080	0.78	0.67	0.74
	Asian	6,658	0.80	0.72	0.79
	Black	8,192	0.65	0.50	0.66
	Hispanic	24,577	0.73	0.58	0.66
	Other	1,536	0.85	0.78	0.85
	Multiple	2,929	0.70	0.57	0.67
	White	21,264	0.77	0.66	0.74
Geometry	Female	44,292	0.75	0.64	0.71
	Male	44,748	0.77	0.67	0.74
	Asian	6,833	0.82	0.75	0.79
	Black	11,143	0.66	0.54	0.64
	Hispanic	29,876	0.72	0.59	0.67
	Other	1,588	0.82	0.76	0.81
	Multiple	3,573	0.73	0.63	0.73
	White	32,025	0.78	0.70	0.76
Integrated Math 1	Female	1,913	0.75	0.63	0.71
	Male	2,087	0.78	0.70	0.76
	Hispanic	2,204	0.71	0.60	0.68
Integrated Math 2	Female	1,467	0.75	0.70	0.77
	Male	1,529	0.78	0.68	0.73
	Hispanic	1,727	0.70	0.62	0.71
Biology/ Life Science	Female	31,883	0.74	0.63	0.69
	Male	32,795	0.76	0.66	0.71
	Asian	3,653	0.79	0.69	0.74
	Black	8,600	0.69	0.54	0.60
	Hispanic	24,969	0.71	0.60	0.66
	Multiple	3,161	0.72	0.58	0.66
	White	20,171	0.78	0.70	0.75

Note. Results for a demographic group were omitted if the sample size for any term was less than 1,000; as such, no data is provided for Integrated Math 3.

8. Scores Align with Their Intended Purpose, Use, and Interpretation

A scale score is a transformation of the examinee ability parameter, θ . Scale scores provide a common numerical metric, but additional information is needed to support score interpretation. *Norm-referenced* information is provided by comparing a student's performance with that of a reference population of students who have taken the test. *Criterion-referenced* information is established by comparing a student's performance with the content domain, as represented by ordered learning statements and predicted proficiency on state summative tests. Assessments may provide norm-referenced, criterion-referenced, or both types of information about student scores as long as there is evidence to support each type of interpretation.

8.1. Scale Scores

MAP Growth course-specific scores are reported on the RIT (**R**asch **u**nIT) scale. RIT scores have a mean of 200, a standard deviation of 10, and a typical range from 100 to 350. During testing, a momentary RIT score is calculated after each response and is used by the adaptive algorithm to select the next item. At the end of the assessment, a final RIT score is calculated using an examinee's entire item response pattern.

8.1.1. Momentary RIT Scores

Each student begins the test with an initial ability estimate based on past test performance. If no prior score is available, a default starting value is assigned according to the test subject and the student's grade. As the test proceeds, each item is selected from a large pool of Rasch-calibrated items based on the student's momentary ability estimate, content requirements, and longitudinal item exposure controls. After each response, the momentary ability estimate is updated using Bayesian methods (Owen, 1975) that incorporate all responses up to that point in the test. The updated momentary ability estimate is factored into the selection of the next item. As this process repeats, each successive estimate is slightly more precise than the previous one. The test continues until the minimum test length has been reached and the standard error associated with the estimate is within acceptable limits or until the maximum test length is reached.

8.1.2. Final RIT Scores

A final ability estimate (i.e., RIT score) is computed via a maximum-likelihood algorithm with fencing (Han, 2016). The score is first estimated on the logit metric and then converted to the RIT scale through a linear transformation.

A limitation of maximum-likelihood estimation in item response theory is that an examinee's ability cannot be estimated when the examinee answers all items correctly or all items incorrectly. Maximum likelihood with fencing (MLEF) addresses this limitation by adding two fictional (fence) items to the response pattern. The lower-fence item is assigned a difficulty of 3.8 logits below the easiest item answered by the examinee and is scored as correct. The upper-fence item is assigned a difficulty of 3.8 logits above the most difficult item completed by the examinee and is scored incorrectly. These fixed item parameters are used to calculate the probability of a correct response to the lower-fence item, $P_{LF}(\theta)$, and the probability of an incorrect response to the upper-fence item, $P_{UF}(\theta)$. The MLEF estimate is obtained by finding the value of θ that maximizes the log-likelihood of the observed item response pattern, $\mathbf{x}$, and the two fence items, as shown in Equation 10.

$$\ell(\theta|\mathbf{x}) = \ln P_{LF}(\theta) + \ln [1 - P_{UF}(\theta)] + \sum_{j=1}^n \{x_j \ln P_j(\theta) + (1 - x_j) \ln [1 - P_j(\theta)]\} \quad (10)$$

This maximization can be completed using numerical optimization methods (e.g., Newton-Raphson). The fenced maximum-likelihood estimate is calculated as shown in Equation 11.

$$\hat{\theta}_{MLEF} = \arg \max_{\theta \in [\theta_L, \theta_U]} \ell(\theta|\mathbf{x}) \quad (11)$$

The final estimate is then transformed to the RIT scale using $RIT = 10\hat{\theta}_{MLEF} + 200$.

8.1.3. Instructional Area Scores

An overall RIT score is calculated from the examinee's full item response pattern. Instructional area scores are calculated in the same way but use only responses to items associated with a specific instructional area. Specifically, the instructional area score is calculated using MLEF with the response pattern limited to a subset of items. The resulting estimates are then converted to the RIT scale using the same linear transformation applied to the overall RIT score.

8.2. MAP Growth Norms

Meaningful interpretation of MAP Growth scores requires appropriate context. Without a reference for comparison, assessment scores are isolated numbers rather than actionable information. MAP Growth achievement and growth norms provide this context by describing the achievement and growth patterns of U.S. public school students and supporting norm-referenced interpretations of individual and group performance.

For MAP Growth course-specific assessments, the most recent available norms are the 2022 MAP Growth norms (NWEA, 2022), which provide reference information that helps educators interpret student achievement and growth relative to a national comparison group. Norm-referenced information can support instructional planning, goal setting, and conversations about student and school performance when used alongside other local evidence.

8.3. Linking Studies and Predicted Proficiency

State summative assessments typically utilize multiple performance levels (e.g., *Below Basic*, *Basic*, *Proficient*, *Advanced*) that are established through a formal standard-setting process. The scores that divide the reporting scale into these performance levels are referred to as cut scores. Because MAP Growth tests are aligned to state standards, MAP Growth RIT scores are often strongly related to performance on corresponding state summative assessments. Linking studies use this relationship to connect MAP Growth scores with state assessment performance levels.

For MAP Growth course-specific assessments, linking studies connect course-specific RIT scores to corresponding state end-of-course or high school summative assessments when sufficient data are available. These studies are typically conducted for courses such as Algebra 1, Geometry, Integrated Math 1, and Biology/Life Science, depending on the structure of the state assessment program and the availability of matched student records.

Linking studies allow educators to use MAP Growth course-specific results to estimate whether students are likely to meet proficiency expectations on corresponding state assessments. This information can help schools identify students who may need additional support in course-specific content before being at risk of not meeting proficiency expectations and provide targeted instruction or intervention before state testing.

8.3.1. Linking Study Methodology

Linking studies use student data from the same spring administration window for MAP Growth course-specific and state summative assessments. Only students who completed both assessments during the period are included. To ensure reliable results, generally at least 1,000 matched student records are required for each course and state assessment. Participation is voluntary, and districts must provide state assessment data and permission to use MAP Growth results. Student records are matched using identifying information such as student name and student ID.

Because not all students in a state take MAP Growth course-specific assessments, the study sample may not fully represent the statewide testing population. To address this limitation, statistical weighting is applied so that the analytic sample more closely reflects the state population in terms of key characteristics such as race, gender, and performance level. These adjustments support a broader interpretation of the linking results for students in the state.

To establish the relationship between MAP Growth course-specific scores and state assessment performance levels, an equipercentile linking method is used. This method identifies scores on the two assessments that correspond to the same percentile rank in their respective score distributions. In practice, a MAP Growth RIT score is identified that corresponds to the same proportion of students meeting or exceeding a given state assessment cut score. This process produces spring MAP Growth RIT cut scores that correspond to state assessment performance-level cut scores.

Once spring RIT cut scores are established, growth information can be used to project corresponding fall and winter cut scores. These projected cut scores allow educators to estimate earlier in the year whether students are on track to meet spring proficiency expectations on the corresponding state assessment.

Overall, this methodology allows schools to use MAP Growth course-specific results throughout the year to evaluate students' progress toward state performance expectations, identify students who may need additional support in course-specific content, and make informed instructional decisions before end-of-course or spring state testing.

8.3.2. Linking Study Accuracy in Predicting Proficiency

The degree to which MAP Growth course-specific tests predict student proficiency status on state summative assessments can be evaluated using classification accuracy statistics from linking studies. These statistics indicate the proportion of students whose MAP Growth scores correctly classify them as proficient or not proficient on the corresponding state summative assessment.

Detailed linking study reports for participating states are available on the [NWEA Connection website](#). Each report describes the study methodology, student sample, and classification accuracy results, including related statistics such as false positive rate, false negative rate, sensitivity, specificity, precision, and area under the curve. Table 8.1 summarizes classification accuracy results for states with available course-specific linking studies.

As shown in Table 8.1, classification accuracy was consistently strong across courses, states, and state summative assessments. Values ranged from 0.80 to 0.89, indicating that MAP Growth course-specific tests correctly classified approximately 80% to 89% of students as proficient or not proficient on the corresponding state assessment. Algebra 1 results ranged

from 0.80 in Texas to 0.87 in Georgia, while Integrated Math 1 in North Carolina had an accuracy of 0.84. Geometry results were similarly strong, with accuracy values of 0.82 in Florida and 0.86 in Ohio. Biology/Life Science results were also high, with accuracy values of 0.89 for Texas Biology and 0.85 for Ohio Biology.

Overall, these findings provide evidence that MAP Growth course-specific tests are strongly aligned with state summative assessments and can support interpretations about whether students are on track to meet state proficiency expectations. The consistency of results across courses and states supports the use of MAP Growth course-specific scores for instructional planning, early identification of students who may need additional support, and data-informed decision-making prior to state summative testing.

Table 8.1. Classification Accuracy for MAP Growth Course-Specific Tests and State Summative Assessments

MAP Growth Course-Specific Test	State	State Summative Assessment	N	Classification Accuracy
Algebra 1	Florida	B.E.S.T. Algebra 1	2,209	0.83
	Georgia	Georgia Milestones Algebra 1	1,109	0.87
	New York	Regents Examination in Algebra I	22,027	0.81
	Ohio	OST Algebra 1	4,448	0.86
	Texas	STAAR Algebra I	3,697	0.80
Integrated Math 1	North Carolina	NC Math 1	937	0.84
Geometry	Florida	B.E.S.T. Geometry	3,342	0.82
	Ohio	OST Geometry	3,820	0.86
Biology	Ohio	OST Biology	1,623	0.85
	Texas	STAAR Biology	2,820	0.89

8.4. Relationships with Measures of the Same Construct

In addition to classification accuracy, linking studies also examine the correlation between MAP Growth course-specific scores and scores on corresponding state summative assessments. These correlations describe the extent to which performance on the two assessments varies together. Higher correlations indicate that students who score higher on MAP Growth course-specific tests also tend to score higher on the state summative assessment.

As shown in Table 8.2, correlations between MAP Growth course-specific tests and state summative assessments were consistently strong across the available linking studies, ranging from 0.70 to 0.89. Algebra 1 correlations ranged from 0.70 in Texas to 0.87 in Ohio, with additional strong relationships observed in Florida, New York, and Georgia. Integrated Math 1 in North Carolina had a correlation of 0.74. Geometry correlations were 0.80 in Florida and 0.89 in Ohio. Biology correlations were also strong, with correlations of 0.84 for both Texas Biology and Ohio Biology.

Overall, these results indicate strong relationships between MAP Growth course-specific scores and state summative assessment scores across courses and states. The findings provide evidence that MAP Growth course-specific tests are meaningful indicators of performance on corresponding state assessments, particularly when interpreted alongside other sources of evidence such as classroom performance and local assessment data.

Table 8.2. Correlations Between MAP Growth Course-Specific Tests and State Summative Assessments

MAP Growth Course-Specific Test	State	State Summative Assessment	N	Correlation
Algebra 1	Florida	B.E.S.T. Algebra 1	2,209	0.71
	Georgia	Georgia Milestones Algebra 1	1,109	0.85
	New York	Regents Examination in Algebra I	22,027	0.73
	Ohio	OST Algebra 1	4,448	0.87
	Texas	STAAR Algebra I	3,697	0.70
Integrated Math 1	North Carolina	NC Math 1	937	0.74
Geometry	Florida	B.E.S.T. Geometry	3,342	0.80
	Ohio	OST Geometry	3,820	0.89
Biology	Ohio	OST Biology	1,623	0.84
	Texas	STAAR Biology	2,820	0.84

8.5. The Learning Continuum

8.5.1. Overview of the Learning Continuum and Its Development

The Learning Continuum is a content exploration tool designed to represent the ordered structure of a MAP Growth assessment. It organizes items in a MAP Growth item pool by difficulty and links them to Learning Statements. Each Learning Statement describes the specific skill, concept, or claim measured by an item, and items measuring the same skill or concept are aligned to the same statement.

Because a MAP Growth item pool contains thousands of items across instructional areas and grade levels, multiple Learning Statements may appear within any given 10-point RIT band. In reports, these statements are grouped by standard or topic to provide a structured view of the content represented across the MAP Growth scale for the item pool of a given test. Through this organization, the Learning Continuum offers a transparent depiction of how assessment content maps onto the RIT scale. Figure 8.1 provides an example of the Learning Continuum tool.

8.5.2. Relationship of the Learning Continuum to the RIT Scale and Student Scores

The Learning Continuum and its Learning Statements represent the content domain underlying the MAP Growth RIT scale by describing the skills and concepts reflected in the item pool at different points along the scale. For MAP Growth course-specific assessments, Learning Statements describe the course-based and prerequisite skills measured by items in the math and Biology/Life Science item pools. Grouping these statements by RIT range illustrates the types of content typically associated with each 10-point band within a course-specific assessment. Importantly, this linkage helps clarify what a student's RIT score represents in relation to the underlying content domain measured by the course-specific test without prescribing instructional placement or defining mastery. Instead, the Learning Continuum provides students, educators, assessment designers, and curriculum specialists with insight into how course-specific items map onto the RIT scale and how student performance relates to specific content features within the assessed course.

8.5.3. Instructional Uses of the Learning Continuum

Educators can use the Learning Continuum to deepen their understanding of the content represented by MAP Growth and to interpret patterns of relative strengths and needs at the classroom or group level. Because it reflects the content of the MAP Growth item pools, the

Learning Continuum helps educators identify skills and concepts students may have encountered during testing and connect MAP Growth data with standards, instructional areas, and curriculum materials.

For course-specific tests, the Learning Continuum may be especially useful for planning support within a defined course, such as identifying prerequisite skills that may need reinforcement or course-level concepts where students may benefit from additional practice, scaffolding, or enrichment. Although MAP Growth course-specific results provide useful information for planning across broader instructional intervals, high-quality formative assessment remains essential for tailoring daily instruction. Used together, course-specific MAP Growth data, the Learning Continuum, and ongoing formative practices provide complementary evidence for instructional decision-making and support student learning within the assessed course content.

Figure 8.1. Learning Continuum Example

171-180	181-190	191-200	201-210	211-220	221-230	231-240	241-250	251-260	261-270	271-280	281-290	291-300
RIT 241-250 ⓘ							RIT 251-260 ⓘ					
Number and Algebraic Methods Describe & Graph Linear Functions, Equations, & Inequalities Write & Solve Linear Functions, Equations, & Inequalities Quadratic and Exponential Functions and Equations							Number and Algebraic Methods Describe & Graph Linear Functions, Equations, & Inequalities Write & Solve Linear Functions, Equations, & Inequalities Quadratic and Exponential Functions and Equations					
Write & Solve Linear Functions, Equations, & Inequalities							Write & Solve Linear Functions, Equations, & Inequalities					
Solve and Evaluate the Reasonableness of Solutions							Solve and Evaluate the Reasonableness of Solutions					
Algebra 1.5.A: solve linear equations in one variable, including those for which the application of the distributive property is necessary and for which variables are included on both sides: - Solves multi-step linear equations with positive and negative rational numbers - Writes and solves a multi-step linear equation in one variable involving a real-world or mathematical context							Algebra 1.5.A: solve linear equations in one variable, including those for which the application of the distributive property is necessary and for which variables are included on both sides: - Solves multi-step linear equations with positive and negative rational numbers - Solves multi-step linear inequalities - Writes and solves a multi-step linear equation in one variable involving a real-world or mathematical context					
Algebra 1.5.B: solve linear inequalities in one variable, including those for which the application of the distributive property is necessary and for which variables are included on both sides; and - Solves multi-step linear inequalities - Solves two-step linear inequalities							Algebra 1.5.B: solve linear inequalities in one variable, including those for which the application of the distributive property is necessary and for which variables are included on both sides; and - Represents the solutions of a multistep linear inequality on a number line - Solves multi-step linear inequalities					
Algebra 1.5.C: solve systems of two linear equations with two variables for mathematical and real-world problems. - Shows the steps to solve a system of linear equations using linear combinations - Solves a system of linear equations algebraically - Writes a system of linear equations given a graph - Writes and solves a system of linear equations involving a real-world or mathematical context							Algebra 1.5.C: solve systems of two linear equations with two variables for mathematical and real-world problems. - Determines the number of solutions to a system of linear equations - Shows the steps to solve a system of linear equations using linear combinations - Shows the steps to solve a system of linear equations using the substitution method - Solves a system of linear equations algebraically					

9. The Test Is Fair for All Examinees

9.1. Test Taker Demographics

The MAP Growth course-specific assessments were administered to large samples of students across courses and testing terms. Sample sizes were largest for Algebra 1, Geometry, Algebra 2, and Biology/Life Science, while the Integrated Math courses had smaller but sufficient sample sizes for demographic summaries. Demographic information is reported for student gender and race/ethnicity because these are the demographic variables most consistently submitted by school districts. Other demographic information, such as English learner status or disability status, is voluntary and not consistently available across districts.

As shown in Table 9.1, gender representation was highly balanced across courses and terms. Male students comprised approximately 50% to 52% of examinees, while female students comprised approximately 48% to 50%. The largest male percentage was observed for Integrated Math 1 in fall at 52.32%, and the lowest was observed for Algebra 2 in winter at 50.29%. Overall, the gender distribution showed minimal variation across fall, winter, and spring administrations, indicating that the course-specific assessment samples were nearly evenly distributed by reported gender.

Table 9.1. Gender Distribution by Course and Term

Course	Term	N	Male (%)	Female (%)
Algebra 1	Fall	233,603	51.65	48.35
	Winter	191,488	51.54	48.46
	Spring	179,232	51.41	48.59
Algebra 2	Fall	112,927	50.46	49.54
	Winter	83,841	50.29	49.71
	Spring	90,791	50.35	49.65
Geometry	Fall	179,524	50.55	49.45
	Winter	141,789	50.42	49.58
	Spring	136,548	50.62	49.38
Integrated Math 1	Fall	13,049	52.32	47.68
	Winter	8,084	51.71	48.29
	Spring	10,445	51.09	48.91
Integrated Math 2	Fall	9,773	51.62	48.38
	Winter	5,998	51.08	48.92
	Spring	8,415	50.91	49.09
Integrated Math 3	Fall	4,888	51.00	49.00
	Winter	3,891	51.25	48.75
	Spring	3,212	50.97	49.03
Biology/Life Science	Fall	116,783	51.17	48.83
	Winter	95,356	51.10	48.90
	Spring	96,352	51.19	48.81

Note. Demographic information was not reported for some examinees and was omitted from these counts.

The racial and ethnic composition of students taking MAP Growth course-specific assessments varied by course but was generally stable across fall, winter, and spring within each course. As shown in Table 9.2, Hispanic students represented the largest group in most courses and terms, particularly in the Integrated Math courses, where they accounted for approximately 36% to 52% of examinees. White students were also strongly represented across courses, ranging from approximately 20% to 37%, with the highest percentages generally observed in Geometry, Algebra 2, and Integrated Math 3. Black students represented approximately 6% to 20% of examinees, with higher percentages in Algebra 1 and Biology/Life Science and lower percentages in Integrated Math 2 and Integrated Math 3. Asian students generally accounted for approximately 5% to 9% of examinees across courses and terms.

Students identifying as American Indian or Alaska Native and Native Hawaiian or Pacific Islander represented smaller proportions of the tested population, generally below 2% and below 0.5%, respectively. Students identifying as Multiple races accounted for approximately 4% to 7% of examinees, while students categorized as Other generally represented about 1% to 8%, with the highest percentage observed for Integrated Math 3 in spring.

Overall, the race/ethnicity distributions were reasonably consistent across testing terms within each course, with some variation in smaller samples, particularly in the Integrated Math courses. These results indicate that the course-specific assessment samples included racially and ethnically diverse groups of students across courses and administrations.

Table 9.2. Race/Ethnicity Distribution by Course and Term

Course	Term	N	Asian (%)	Black (%)	NH/PI (%)	AI/AN (%)	Hispanic (%)	White (%)	Multiple (%)	Other (%)
Algebra 1	Fall	233,603	5.76	19.37	0.25	1.36	34.81	31.58	4.34	2.53
	Winter	191,488	6.17	20.13	0.24	1.40	34.33	31.44	4.19	2.10
	Spring	179,232	6.15	18.08	0.23	1.27	35.74	30.84	5.00	2.70
Algebra 2	Fall	112,927	6.54	12.10	0.16	1.43	36.61	34.94	4.72	3.50
	Winter	83,841	6.15	13.31	0.14	1.62	35.25	35.69	4.93	2.89
	Spring	90,791	6.13	12.30	0.15	1.17	38.49	32.00	6.43	3.33
Geometry	Fall	179,524	6.91	14.96	0.21	1.27	33.03	36.77	4.43	2.42
	Winter	141,789	7.34	15.59	0.20	1.25	32.71	36.72	4.29	1.90
	Spring	136,548	7.08	13.89	0.19	1.05	34.52	35.51	5.39	2.38
Integrated Math 1	Fall	13,049	5.32	12.76	0.29	0.92	45.43	26.24	5.13	3.92
	Winter	8,084	7.71	14.20	0.21	1.18	48.19	20.27	4.38	3.86
	Spring	10,445	6.00	9.27	0.33	0.75	50.61	23.54	4.55	4.96
Integrated Math 2	Fall	9,773	7.90	6.12	0.25	0.57	45.47	30.61	5.57	3.51
	Winter	5,998	9.34	8.44	0.22	0.68	52.35	21.02	3.98	3.97
	Spring	8,415	7.26	7.97	0.27	0.40	48.31	26.44	4.98	4.36
Integrated Math 3	Fall	4,888	7.79	8.41	0.39	0.51	35.78	34.12	5.26	7.73
	Winter	3,891	9.20	8.35	0.15	0.72	37.37	33.49	4.37	6.35
	Spring	3,212	7.78	7.13	0.19	0.31	43.59	25.78	4.20	11.02
Biology/ Life Science	Fall	116,783	4.67	14.80	0.18	1.71	39.34	31.18	5.33	2.79
	Winter	95,356	5.20	15.42	0.19	1.47	38.48	31.50	5.38	2.35
	Spring	96,352	4.67	14.40	0.21	1.38	40.22	30.98	6.88	1.26

Note. Examinees missing demographic information are omitted from this analysis. NH/PI = Native Hawaiian or Pacific Islander; AI/AN = American Indian or Alaska Native

9.2. Universal Design

Test development incorporates Universal Design for Learning (UDL) principles to address the needs of diverse populations of students taking the MAP Growth assessments. The NWEA content team applies the UDL principles summarized in Table 9.3 (Thompson et al., 2002) and UDL guidelines (CAST, 2018) when creating test items. These principles improve tests and test fairness by removing characteristics that are unrelated to the measured construct but may inadvertently affect test scores. The result is a more accurate score for the student and a clearer picture of what the student knows and can do. It also provides a framework for incorporating flexibility in the ways content is presented and how students respond or show their knowledge. It also allows multiple ways for students to be engaged.

Table 9.3. Universal Design for Learning Principles

UDL Principle	Description
Inclusive assessment population	Field tests should include students with a wide range of abilities, students with limited English proficiency, and students across racial, ethnic, and socioeconomic lines.
Precisely defined constructs	The test design is clear on the construct(s) to be measured, the purpose for which scores will be used, and inferences that will be made from the scores. Universally designed assessments do this by removing barriers, which is referred to as construct-irrelevant variance.
Accessible, non-biased items	To ensure the quality of items, a differential item functioning (DIF) analysis can investigate whether certain items perform differently for various subpopulations. Additionally, panels of bias, sensitivity, fairness, and accessibility experts have contributed to the review of items and the production of extensive guidelines and resources guiding item development.
Amenable to accommodations	Accommodations are used to increase access to assessments and to the items within the assessments. Accommodations change the environment in which the test is presented or responded to and are typically used by students with disabilities and by English language learners (ELLs).
Simple, clear, and intuitive instructions and procedures	Assessments should be easy to understand regardless of a student's knowledge and experience. The instructions and procedures of the test and the items themselves should not create barriers for students. The student must be able to access the test as intended.
Maximum readability and comprehensibility	Ensuring readability and comprehensibility is important for clarity and access purposes. It is vital that the construct to be measured is presented clearly with plain language and at the appropriate reading level.
Maximum legibility	This refers to the capability of being deciphered with ease.

9.3. Accommodations

MAP Growth utilizes several features to improve test fairness and provide more precise and valid assessment measurement. These features fall within three categories: (1) universal features, (2) designated features, and (3) accommodations.

Local schools and districts may determine whether certain features are considered universal, designated, or accommodations. Schools and districts are encouraged to follow their current state accessibility and accommodation guidelines when deciding which features are appropriate for an individual student. The policy at NWEA is aligned with the *CCSSO Accessibility Manual* (CCSSO, 2016). The goal is to provide a universal approach and make the use of features and accommodations as easy as possible for students and educators.

9.3.1. Universal Features

Table 9.4 presents the universal features available for MAP Growth. Universal features are accessibility supports that are available to all students as they access instructional or assessment content. They are either embedded and provided digitally through instructional or assessment technology (such as a keyboard) or non-embedded and provided non-digitally at the local level (such as scratch paper).

Table 9.4. Available Universal Features

Feature	Description
Embedded	
Amplifications	A student raises or lowers the volume control, as needed, using headphones.
Calculator	A student can access an on-screen digital calculator for calculator-allowed items. If the calculator is not appropriate (e.g., for a student who is blind), the student may use a calculator provided with assistive technology (such as a talking calculator or a braille calculator).
Highlighter	A student can use this digital feature to mark desired text, items, or response options with a color.
Zoom (item-level)	The student can enlarge the size of text and graphics on a given screen. This feature allows students to view material in magnified form on an as-needed basis. The student may enlarge test content at least fourfold. The system allows magnifying features to work in conjunction with other accessibility features and accommodations provided.
Line reader	A student can use this tool as a guide when reading text.
Answer choice eliminator	A student can cross out answer choices that do not appear to be correct.
Notepad	A student can make notes or record responses virtually.
Math tools	These digital tools (e.g., ruler, protractor, calculator) are used for tasks related to math items. They are available only with the specific items for which one or more of these tools would be appropriate.
Keyboard navigation (keyboard shortcuts, two-switch system)	A student can navigate through test content by using the keyboard (e.g., the arrow keys). This feature may differ depending on the testing device.
Non-Embedded	
Breaks (frequent breaks)	A student can take breaks, when needed, to reduce cognitive fatigue.

Feature	Description
English dictionary	A student can use an English dictionary.
Noise buffer (headphones, audio aids)	A student can use noise buffers to minimize distractions or filter external noises during testing. Noise buffers must be compatible with the requirements of the test.
Scratch paper (blank paper)	A student can use scratch paper or an individual erasable whiteboard to make notes or record responses. The school must also provide a marker, pen, or pencil. All scratch paper must be collected and securely destroyed at the end of each test to maintain test security. The student can use an assistive technology device to take notes instead of using scratch paper if the device is approved by the state. Test administrators must ensure that all notes taken on an assistive technology device are deleted after the test.
Spanish dictionary	A student can use a Spanish dictionary, if necessary.
Thesaurus	A student can use a thesaurus containing synonyms of terms.

9.3.2. Designated Features

Table 9.5 presents the designated features available for MAP Growth. Designated features are available when an educator (or team of educators, including the parents/guardians and the student, if appropriate) indicates that there is a need for them. Designated features must be assigned to a student by trained educators and then enabled in the system by the proctor. Embedded designated features such as text-to-speech (TTS) are provided digitally through instructional or assessment technology. Non-embedded designated features (such as a magnification device) are provided locally.

Table 9.5. Available Designated Features

Feature	Description
Embedded	
Text-to-speech (TTS) (audio support, spoken audio)	A student can hear computer-generated audio of the item content.
Non-Embedded	
Bilingual dictionary (word-to-word dictionary in English and native language)	A student can use a bilingual/dual language word-to-word dictionary as a language support.
Color contrast	A student can display the test content of online items in different colors.
Human reader (human read aloud, read aloud)	A qualified human reader can read the test and item content aloud.
Magnification device (low-vision aids)	A student can adjust the size of specific areas of the screen (e.g., text, formulas, tables, and graphics) with an assistive technology device. Magnification allows the student to increase the size to a level that is not provided by the zoom universal feature.
Native language translation	A test administrator who is fluent in the student's native language can translate test and item content.

Feature	Description
Separate setting (alternate location)	A school can alter a test location so that the student is tested in a setting that's different from what's available for most students.
Student reads test aloud	A student can read the test content aloud. This feature must be administered in a one-on-one test setting.

9.3.3. Accommodations

Table 9.6 presents the accommodations available for MAP Growth. Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations are provided digitally through instructional or assessment technology. Non-embedded accommodations (such as a scribe) are provided locally. Accommodations are generally available to students for whom there is a documented need on an Individualized Education Program (IEP) or 504 accommodation plan, although some states also offer accommodations for ELLs.

Table 9.6. Available Accommodations

Accommodation	Description
Embedded	
Text-to-speech (TTS) (audio support, spoken audio)	A student can hear computer-generated audio of the item content.
Non-Embedded	
Abacus (individual manipulatives)	May be used in place of scratch paper for students who typically use an abacus.
Assistive technology (alternate response options, word processor, or similar keyboarding device to respond to items)	A student can use assistive technology, which includes supports such as typing on customized keyboards; assistance with using a mouse, mouth or head stick, or other pointing devices; sticky keys; touch screen; and trackball.
Calculator (calculation device)	A student can use a specific calculation device (e.g., large key, talking, or other).
Extended time	Schools can allow flexible scheduling for a student test administration (e.g., testing longer than a scheduled test session, multiple breaks)
Human signer (sign language, sign interpretation of test)	A test administrator who is fluent in the language can sign test and item content. The student may also dictate responses by signing.
Multiplication table	A student can use a paper-based single digit (1–9) multiplication table.
Refreshable braille	A student can use a refreshable braille device that provides a raised-dot code that they can read with their fingertips.
Screen reader	A student with no or low vision can use a software application that identifies and interprets what is being displayed on the screen (e.g., text, images).

Accommodation	Description
Scribe	A student can dictate their responses to an experienced educator who records verbatim what the student dictates.

9.3.4. Third-Party Assistive Software

Third-party software features, such as those in Table 9.7, are allowed when not using the lockdown browser. If students try using these tools with the lockdown browser, they will have limited or no functionality. Therefore, NWEA recommends that students who need to use specific features use browser-based testing. If students use the lockdown browsers, NWEA recommends they launch the third-party tool prior to launching the lockdown browser.

Table 9.7. Third-Party Assistive Software

Third-Party Software	Description
ZoomText	A powerful computer access solution designed for the visually impaired. It offers a combination of magnification and reading tools, as well as enhancements to colors, pointers, and cursors. It works for both Mac® and Windows® operating systems.
Chromebook® magnification	Chromebook has a built-in screen magnifier. This allows users to zoom in and out anywhere on the screen.
Windows magnifier	The magnifier in Windows is part of the Ease of Access Center and can be used to enlarge different parts of the screen. Users of Windows versions 7 and up can choose from either full screen or lens magnification modes.
Zoom on Mac and iPad®	Mac computers and iPads have a built-in screen magnifier that can magnify a screen up to 40 times its normal display size.
Chromebook color contrast	High-contrast mode inverts the picture so that a white background appears black, black text appears white, and colors are inverted (for example, blue text or graphics become orange).
Windows color contrast	Windows supports high-contrast themes for the OS and apps that users may choose to enable. High-contrast themes use a small palette of contrasting colors that makes the interface easier to see.
Mac and iPad color contrast	Increase the readability of the screen on your MacBook® or iPad by increasing the contrast of the display. Increase the contrast of the whole screen or emphasize borders between items in the Display section of the Accessibility settings.
JAWS	Job Access with Speech (JAWS) is the world's most popular screen reader, developed for computer users whose vision loss prevents them from seeing screen content or navigating with a mouse. JAWS provides speech and braille output for the most popular computer applications.
Refreshable braille device	A refreshable braille device provides a raised-dot code that individuals read with their fingertips.

9.4. Differential Item Functioning

Differential item functioning (DIF) analysis is conducted to evaluate whether individual test items function equivalently for different groups of examinees who have the same underlying level of

achievement. An item exhibits DIF when students from different groups, matched on overall ability, have different probabilities of responding correctly to that item. DIF analyses are a key component of test fairness and validity evaluations, as they help identify items that may advantage or disadvantage particular groups for reasons unrelated to the construct being measured. In the context of MAP Growth, DIF analyses are used to monitor item performance across gender and racial/ethnic groups and to ensure that score interpretations are appropriate and comparable for all students.

9.4.1. Mantel-Haenszel Procedure and ETS DIF Classification Levels

DIF in MAP Growth is evaluated using the Mantel–Haenszel (MH) procedure, a widely used statistical method for detecting uniform DIF in dichotomously scored test items. The MH procedure compares the odds of correct responses for a focal group and a reference group while conditioning on overall test performance, which serves as a proxy for the latent trait being measured (Mantel & Haenszel, 1959; Holland & Thayer, 1988). Examinees are first grouped into score strata based on their total test score or estimated ability. Within each stratum, the odds of a correct response for the focal group are compared with those for the reference group. These stratum-specific odds ratios are then combined into a common odds ratio that summarizes the degree and direction of DIF across the score distribution.

Specifically, examinees are divided into k score strata based on their RIT score. Within each stratum, a 2 x 2 contingency table is created such as the one shown.

	Correct	Incorrect
Focal	A_k	B_k
Reference	C_k	D_k

The stratum-specific odds ratio is given by Equation 12.

$$OR_k = \frac{A_k D_k}{B_k C_k} \quad (12)$$

The Mantel-Haenszel common odds ratio aggregates the odds ratios across strata as shown in Equation 13.

$$\hat{\alpha}_{MH} = \frac{\sum_{k=1}^K \frac{A_k D_k}{N_k}}{\sum_{k=1}^K \frac{B_k C_k}{N_k}} \quad (13)$$

where $N_k = A_k + B_k + C_k + D_k$.

For interpretability, the MH odds ratio is typically transformed to the ETS delta metric (Δ_{MH}), which places DIF effects on a standardized scale centered at zero. Positive values indicate DIF favoring the focal group, whereas negative values indicate DIF favoring the reference group. The common-odds ratio is converted to the ETS Delta scale using Equation 14.

$$\Delta_{MH} = -2.35 \ln(\hat{\alpha}_{MH}) \quad (14)$$

which rescales the values to be centered at zero.

Statistical significance is evaluated using a chi-square test associated with the MH statistic, but practical significance is emphasized through effect-size classification rather than reliance on significance testing alone, particularly in large samples (Holland & Thayer, 1988; Dorans & Holland, 1993). Statistical significance is assessed using the Mantel-Haenszel chi-square statistic shown in Equation 15.

$$\chi^2 = \frac{(\sum_k |A_k - E(A_k)| - 0.5)^2}{Var(A_k)} \quad (15)$$

where $E(A_k)$ and $Var(A_k)$ are the expected value and variance of A_k under the null hypothesis of no DIF. Because large sample sizes can yield statistically significant results for trivial effects, practical significance is emphasized in effect size based DIF classification.

ETS DIF classification guidelines are used to categorize items based on the magnitude of the Δ_{MH} statistic. Items are classified into three levels, as outlined in Table 9.8: Category A (negligible DIF) includes items with small DIF effects that are considered inconsequential for score interpretation; Category B (moderate DIF) includes items with intermediate DIF effects that may warrant review, especially if they occur systematically across forms or administrations; Category C (large DIF) includes items with substantial DIF effects that are likely to be meaningful and typically require content review and potential revision or removal. Each category is further divided by directionality (favoring the focal or reference group), allowing for more detailed interpretation of DIF patterns (Dorans & Kulick, 1986; Zieky, 1993).

Table 9.8. DIF Classification Levels

DIF Classification	Criteria	Interpretation
A	Not statistically significant or $ \Delta_{MH} < 1.0$	Negligible DIF
B	Statistically significant and $ \Delta_{MH} < 1.5$	Moderate DIF
C	Statistically significant and $ \Delta_{MH} \geq 1.5$	Large DIF

Together, the Mantel–Haenszel procedure and ETS DIF classification framework provide a rigorous and well-established approach for monitoring item fairness. Their use in MAP Growth supports the ongoing evaluation of item performance and helps ensure that test scores are comparable and valid across demographic groups.

9.4.2. MAP Growth DIF Results

DIF analyses were conducted for MAP Growth course-specific assessments across fall, winter, and spring administrations. DIF was evaluated by comparing focal and reference groups defined by gender (Female vs. Male) and race/ethnicity (Asian, Black, and Hispanic students, each compared with White students). Items were classified into DIF categories (A, B, or C) with directionality indicating whether DIF favored the focal (+) or reference (–) group.

Across courses, terms, and group comparisons, the large majority of items were classified as A-level, indicating negligible DIF. For Algebra 1, Algebra 2 and Geometry, most comparisons had more than 90% of items classified as A-level when A+ and A– were combined. A-level percentages were especially high for Hispanic-White comparisons in these courses, often exceeding 95%. Biology/Life Science showed a similar pattern, with most comparisons also falling above 90% A-level DIF across terms. These results indicate that most items functioned comparably across focal and reference groups and did not exhibit meaningful DIF.

Integrated Math 1, Integrated Math 2, and Integrated Math 3 also showed predominantly A-level classifications, but the results for these courses should be interpreted with additional caution because the number of eligible items was substantially smaller, particularly for some race/ethnicity comparisons. For example, several Integrated Math 2 and Integrated Math 3 comparisons included fewer than 100 eligible items, and some Integrated Math 3 spring race/ethnicity comparisons had no eligible items after applying the minimum response criterion. Despite these smaller item counts, the available results generally showed that most eligible items were classified as A-level.

B-level DIF, indicating moderate DIF, accounted for a relatively small proportion of items across most courses and comparisons. For Algebra 1, Algebra 2, and Geometry, B-level percentages were generally below 10% when B+ and B- were combined, although some Asian-White comparisons in Algebra 1 and Algebra 2 showed somewhat higher B-level percentages. Biology/Life Science showed similarly low to moderate B-level rates, with most comparisons below approximately 7%. B-level DIF appeared in both directions, favoring either the focal or reference group depending on the course, term, and comparison, and there was no consistent pattern indicating systematic DIF against a particular group.

C-level DIF, indicating large DIF, was uncommon overall. For Algebra 1 and Geometry, most comparisons had combined C-level percentages below approximately 4%, with a small number of exceptions. Algebra 2 showed somewhat higher C-level percentages for Asian-White comparisons, particularly in winter and spring. Biology/Life Science generally had low C-level percentages, although Black-White comparisons in winter and spring showed slightly elevated values relative to other Biology/Life Science comparisons. In the Integrated Math courses, some C-level percentages appeared larger in certain comparisons, but these cases were often based on small numbers of eligible items and should be interpreted cautiously. Overall, the low prevalence of C-level DIF supports the conclusion that very few items showed substantial differential functioning.

A summary of the DIF classification results is presented in Table 9.9.

Table 9.9. DIF Classification Summary by Course, Term, and Group

Course	Term	Focal	Reference	N Items	Percent in DIF Class					
					A+	A-	B+	B-	C+	C-
Algebra 1	Fall	Female	Male	1,749	49.11	43.85	2.40	3.32	0.11	1.20
		Black	White	1,199	46.87	49.12	0.92	2.67	0.08	0.33
		Asian	White	964	42.01	42.95	5.08	6.22	1.56	2.18
		Hispanic	White	1,361	44.01	52.98	1.62	1.10	0.07	0.22
	Winter	Female	Male	1,761	51.33	41.00	2.44	3.80	0.17	1.25
		Black	White	1,097	45.31	48.22	2.28	3.56	0.09	0.55
		Asian	White	824	41.99	45.02	4.25	4.98	1.82	1.82
		Hispanic	White	1,292	43.42	52.94	1.55	1.86	0.00	0.23
	Spring	Female	Male	1,868	54.23	39.67	2.62	2.84	0.00	0.64
		Black	White	1,093	45.93	45.93	2.47	4.48	0.64	0.55
		Asian	White	720	41.81	43.61	5.00	4.17	3.33	2.08
		Hispanic	White	1,373	44.06	51.49	1.82	2.18	0.22	0.22
Algebra 2	Fall	Female	Male	1,205	53.36	38.34	2.32	4.15	0.17	1.66
		Black	White	620	52.26	40.65	3.23	3.06	0.16	0.65

Course	Term	Focal	Reference	N Items	Percent in DIF Class					
					A+	A-	B+	B-	C+	C-
		Asian	White	641	46.33	32.14	10.92	3.28	4.99	2.34
		Hispanic	White	759	54.41	43.21	0.79	1.45	0.13	0.00
		Female	Male	1,129	56.60	34.90	3.01	3.72	0.27	1.51
		Black	White	519	49.71	40.08	4.43	3.85	1.35	0.58
	Winter	Asian	White	495	43.43	34.55	9.90	4.65	5.25	2.22
		Hispanic	White	754	46.82	49.07	2.25	1.59	0.00	0.27
		Female	Male	1,254	57.81	36.20	1.91	2.39	0.08	1.59
		Black	White	552	49.64	40.58	3.44	3.44	2.36	0.54
	Spring	Asian	White	481	43.45	34.10	8.94	5.20	6.65	1.66
		Hispanic	White	895	48.94	46.37	2.35	1.90	0.45	0.00
Geometry	Fall	Female	Male	1,301	51.58	41.81	1.69	3.46	0.00	1.46
		Black	White	964	48.03	48.24	1.56	1.76	0.21	0.21
		Asian	White	957	44.41	46.81	3.76	3.34	0.63	1.04
		Hispanic	White	1,120	43.39	54.82	0.45	1.34	0.00	0.00
	Winter	Female	Male	1,391	51.55	42.34	1.94	3.02	0.14	1.01
		Black	White	932	44.74	47.53	3.33	3.11	0.75	0.54
		Asian	White	885	43.05	48.02	2.82	3.95	0.23	1.92
		Hispanic	White	1,100	42.64	55.45	0.45	1.00	0.27	0.18
	Spring	Female	Male	1,577	52.69	41.22	2.28	2.79	0.06	0.95
		Black	White	1,023	44.97	41.25	5.67	4.11	2.35	1.66
		Asian	White	896	41.74	44.87	5.92	4.69	1.67	1.12
		Hispanic	White	1,339	48.69	45.71	2.46	1.57	0.82	0.75
Integrated Math 1	Fall	Female	Male	268	49.25	42.16	2.61	1.87	1.12	2.99
		Black	White	148	45.95	50.00	0.68	0.00	1.35	2.03
		Asian	White	145	48.97	44.14	2.07	1.38	0.00	3.45
		Hispanic	White	212	50.94	45.75	0.47	1.89	0.00	0.94
	Winter	Female	Male	245	50.20	42.86	2.86	2.86	0.41	0.82
		Black	White	67	38.81	49.25	0.00	5.97	0.00	5.97
		Asian	White	40	45.00	55.00	0.00	0.00	0.00	0.00
		Hispanic	White	198	51.01	43.43	2.53	2.02	0.00	1.01
	Spring	Female	Male	359	47.63	41.50	5.57	3.90	0.28	1.11
		Black	White	128	47.66	47.66	0.78	2.34	0.00	1.56
		Asian	White	121	45.45	45.45	1.65	2.48	1.65	3.31
		Hispanic	White	301	48.17	46.18	2.66	1.33	0.66	1.00
Integrated Math 2	Fall	Female	Male	278	47.84	41.01	3.60	4.32	0.72	2.52
		Black	White	134	40.30	53.73	1.49	0.75	2.99	0.75
		Asian	White	147	44.90	48.30	1.36	3.40	2.04	0.00
		Hispanic	White	220	48.18	49.09	0.45	1.82	0.00	0.45
	Winter	Female	Male	210	46.67	46.19	1.90	3.33	0.00	1.90
		Black	White	20	45.00	50.00	0.00	0.00	0.00	5.00
		Asian	White	15	26.67	73.33	0.00	0.00	0.00	0.00

Course	Term	Focal	Reference	N Items	Percent in DIF Class					
					A+	A-	B+	B-	C+	C-
	Spring	Hispanic	White	162	50.62	43.83	0.62	3.09	0.62	1.23
		Female	Male	302	49.01	41.06	1.66	4.30	0.66	3.31
		Black	White	91	39.56	51.65	1.10	3.30	3.30	1.10
		Asian	White	88	47.73	45.45	0.00	0.00	2.27	4.55
		Hispanic	White	248	41.13	51.61	3.23	2.82	0.40	0.81
Integrated Math 3	Fall	Female	Male	153	41.18	45.75	4.58	5.88	1.31	1.31
		Black	White	49	32.65	65.31	0.00	0.00	0.00	2.04
		Asian	White	40	60.00	32.50	2.50	0.00	0.00	5.00
		Hispanic	White	115	34.78	54.78	1.74	4.35	1.74	2.61
	Winter	Female	Male	142	43.66	45.77	5.63	3.52	0.00	1.41
		Black	White	13	53.85	46.15	0.00	0.00	0.00	0.00
		Asian	White	9	55.56	44.44	0.00	0.00	0.00	0.00
		Hispanic	White	93	45.16	45.16	4.30	2.15	1.08	2.15
	Spring	Female	Male	84	45.24	44.05	4.76	4.76	1.19	0.00
		Black	White	0	—	—	—	—	—	—
		Asian	White	0	—	—	—	—	—	—
		Hispanic	White	35	48.57	51.43	0.00	0.00	0.00	0.00
Biology/ Life Science	Fall	Female	Male	652	48.01	46.32	1.69	3.37	0.00	0.61
		Black	White	536	50.75	45.15	0.75	2.80	0.37	0.19
		Asian	White	482	51.04	41.49	2.49	3.53	0.62	0.83
		Hispanic	White	551	47.55	50.82	0.54	0.54	0.00	0.54
	Winter	Female	Male	681	50.81	45.52	1.17	1.76	0.00	0.73
		Black	White	570	48.25	43.16	2.46	2.81	1.75	1.58
		Asian	White	499	45.09	46.09	3.61	3.61	0.80	0.80
		Hispanic	White	595	42.86	52.77	1.51	1.34	0.67	0.84
	Spring	Female	Male	663	54.00	41.93	0.90	2.87	0.15	0.15
		Black	White	568	44.89	47.01	1.58	2.99	2.29	1.23
		Asian	White	492	42.07	50.61	2.85	2.44	1.42	0.61
		Hispanic	White	568	44.19	52.99	1.06	0.70	0.53	0.53

Note. DIF analyses included items with at least 1,000 usable responses within each course, term, and focal/reference comparison, except for Integrated Math 1, Integrated Math 2, and Integrated Math 3, for which a minimum of 300 usable responses was used due to lower item-level response volume.

10. Score Reports Facilitate the Interpretation and Use of Scores

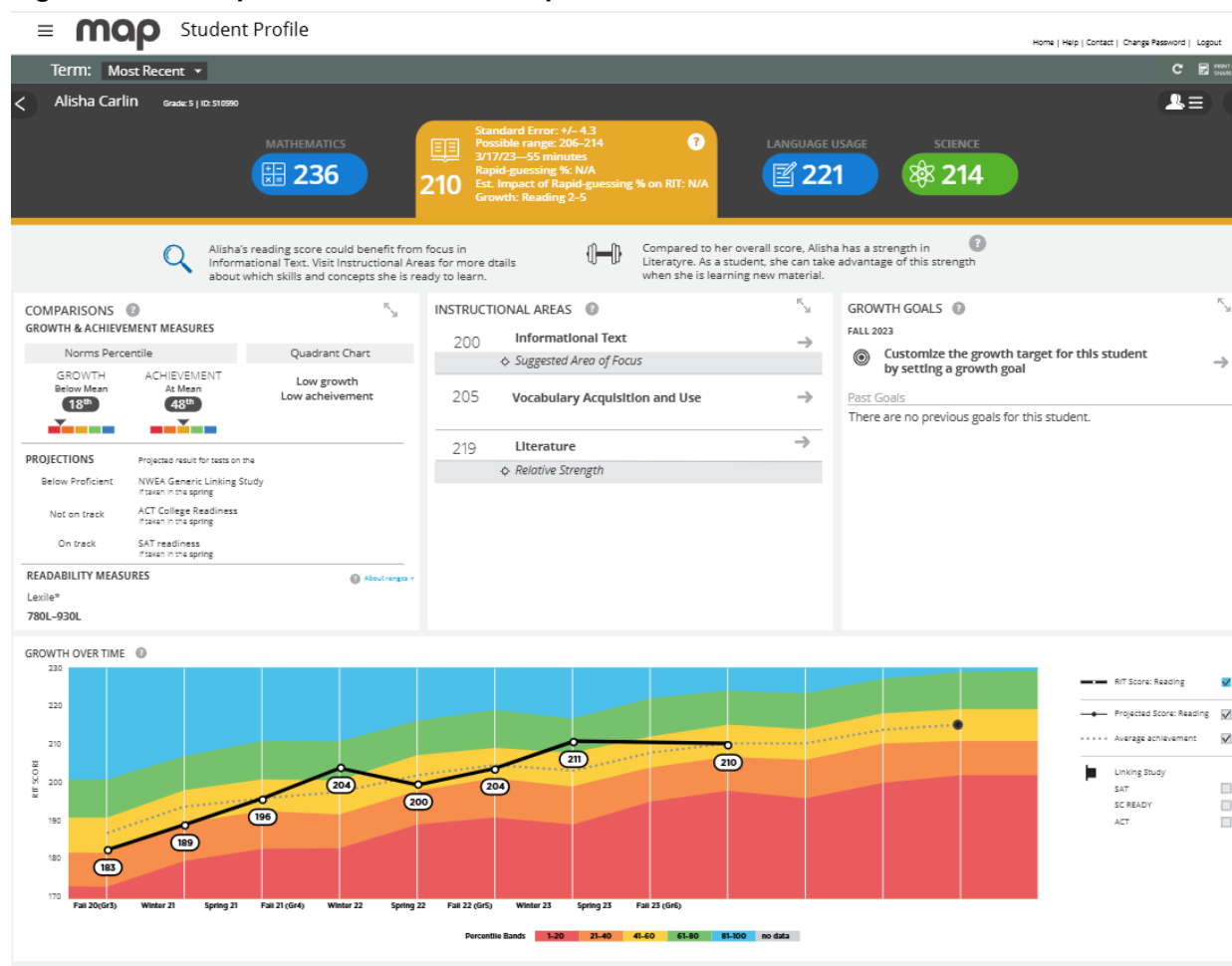
MAP Growth reports facilitate score interpretation by clearly presenting assessment information to a variety of users. Reports are designed to support inferences at the student level and at higher levels of aggregation, such as classroom, school, and district. Each report presents key information to support decision-making at its intended level.

10.1. Student Profile Report

The Student Profile report is NWEA's most comprehensive student-level report. It uses MAP Growth assessment data to create an individualized learner profile that allows educators, caregivers, and students to monitor performance within and across school years. For MAP Growth course-specific assessments, the report provides information about student performance in the assessed course, including current and prior overall RIT scores, instructional area scores, growth information (when available), longitudinal trends, and percentile comparisons. Educators can drill down, customize, and adjust displays from the online dashboard. Additional resources, including a video about how educators can use the Student Profile, are provided at <https://teach.mapnwea.org/impl/maphelp/Content/Data/SampleReports/StudentProfile.htm>.

With this intuitive report, educators can understand how a student is performing, support instructional planning, and set goals collaboratively with the student and caregivers. Norms data, when available for the assessment, provide context for interpreting growth between testing periods and comparing a student's performance with similar students nationwide. The report may also provide proficiency projections for corresponding state summative assessments when linking study results are available. An example Student Profile report is presented in Figure 10.1.

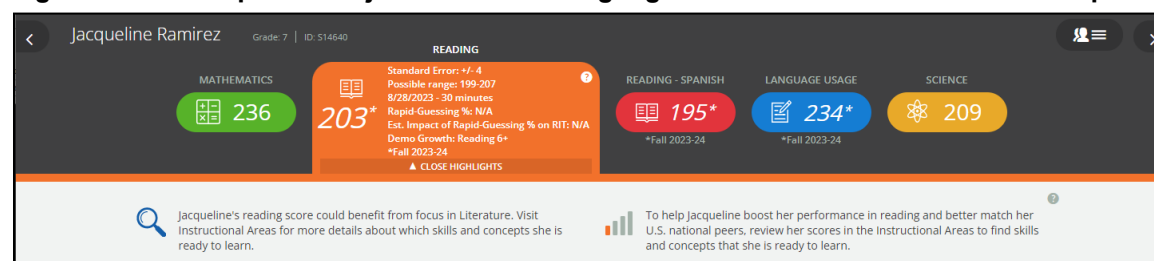
Figure 10.1. Example of Student Profile Report



At the top of the Student Profile, the Subject Scores and Highlights sections (spotlighted in Figure 10.2) provide high-level information about each student's results:

- The Subject Scores section presents the overall RIT score in each assessed subject, along with key details that help contextualize the result. Tabs are color-coded by achievement quintile.
- The Highlights section provides a narrative summary of the student's results, including areas of strength and opportunity.

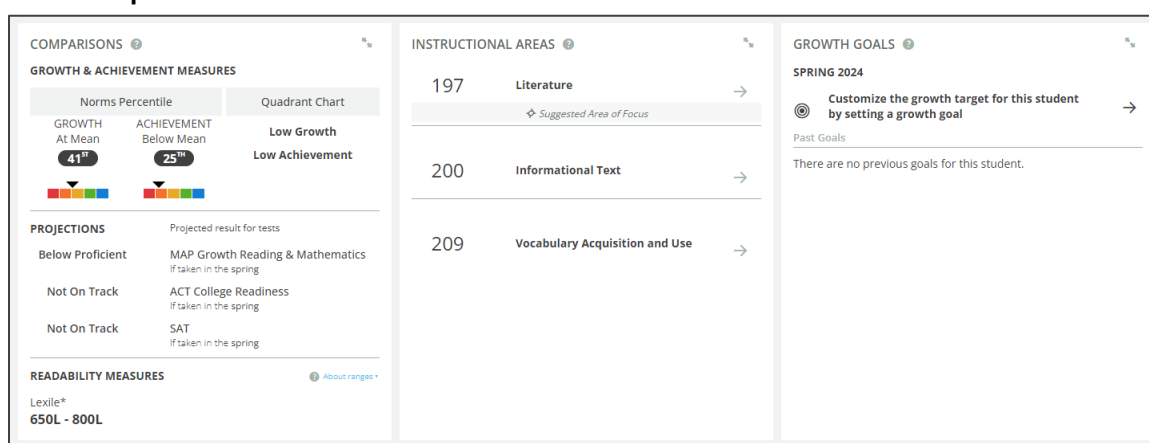
Figure 10.2. Example of Subject Scores and Highlights Sections of Student Profile Report



The middle section of the Student Profile report, highlighted in Figure 10.3, presents more detailed information in the Comparisons, Instructional Areas, and Growth Goals sections:

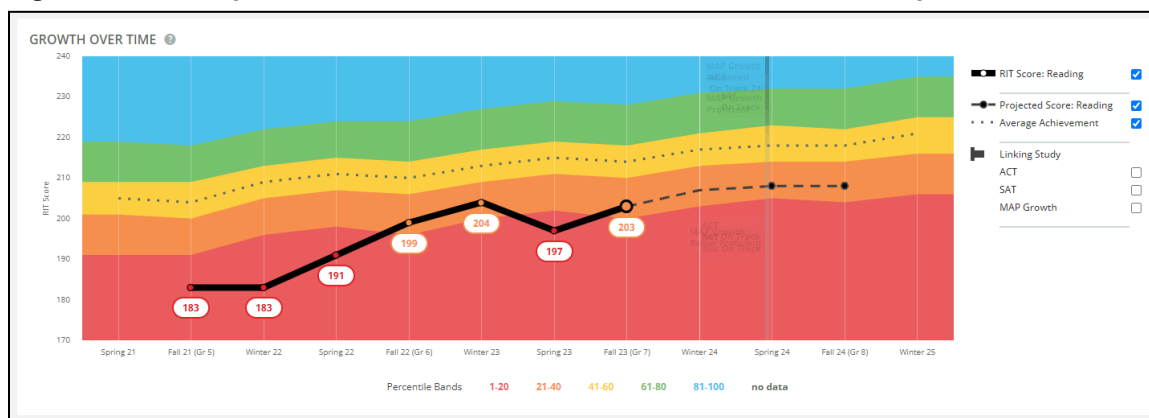
- The Comparisons section places MAP Growth scores into a broader interpretive context. Information may include growth and achievement percentiles, projected proficiency on state summative assessments and college readiness exams, and Lexile or Quantile measures.
- The Instructional Areas section presents scores for each instructional area. Lower scores appear near the top to help educators identify areas for additional support, while higher scores appear near the bottom to highlight areas of relative strength.
- The Growth Goals section allows educators and students to set a growth or performance target for the student.

Figure 10.3. Example of Comparisons, Instructional Areas, and Growth Goals Sections of Student Profile Report



At the bottom of the Student Profile report, the Growth Over Time section, shown in Figure 10.4, displays student growth over time based on current and previous assessment results. It also provides predictive information about typical growth and allows educators to enter specific growth or performance goals for a student.

Figure 10.4. Example of Growth Over Time Section of Student Profile Report

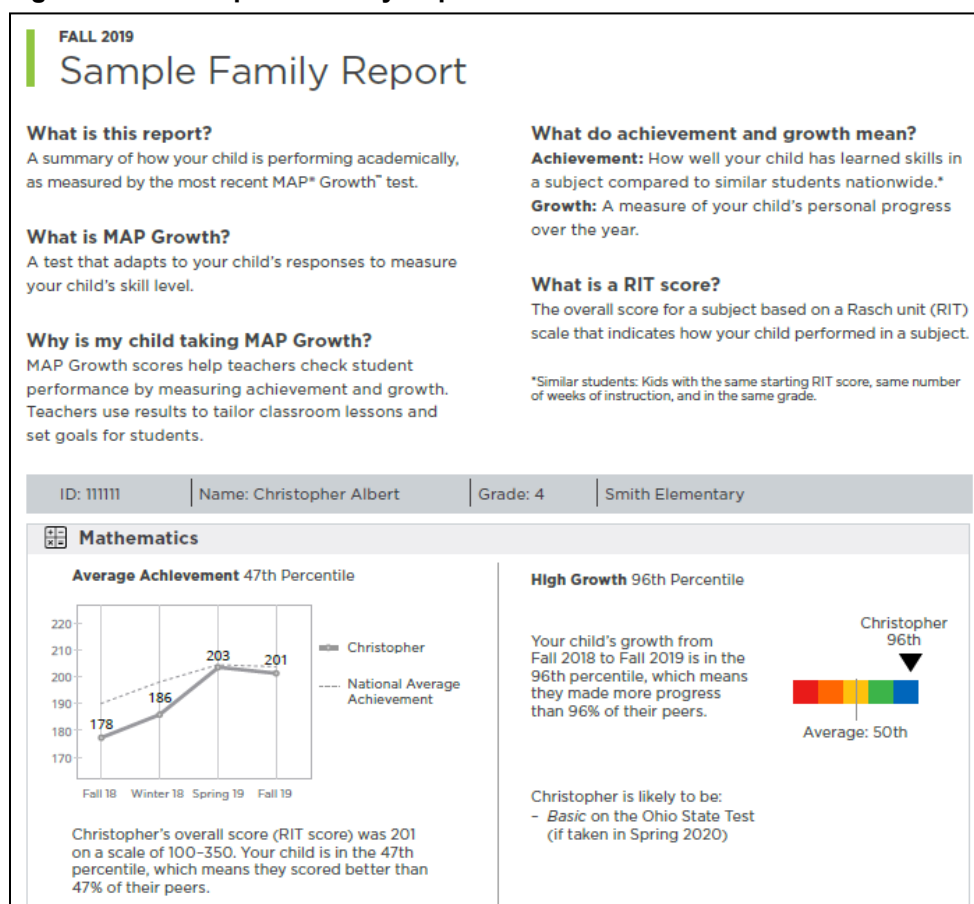


10.2. Family Report

The Family report is one of many MAP Growth reports that can be used to guide conversations with students and their families or caregivers. Educators can use this report to share a clear summary of how each student is performing academically, as measured by MAP Growth.

As shown in the example presented in Figure 10.5, the Family report helps families and caregivers understand a student's performance in the assessed subject, including areas where the student may benefit from additional support and how the student's performance compares with similar students nationwide. The report also includes questions that educators, families, caregivers, and students can use to support meaningful conversations about assessments during conferences.

Figure 10.5. Example of Family Report



10.3. Class Reports

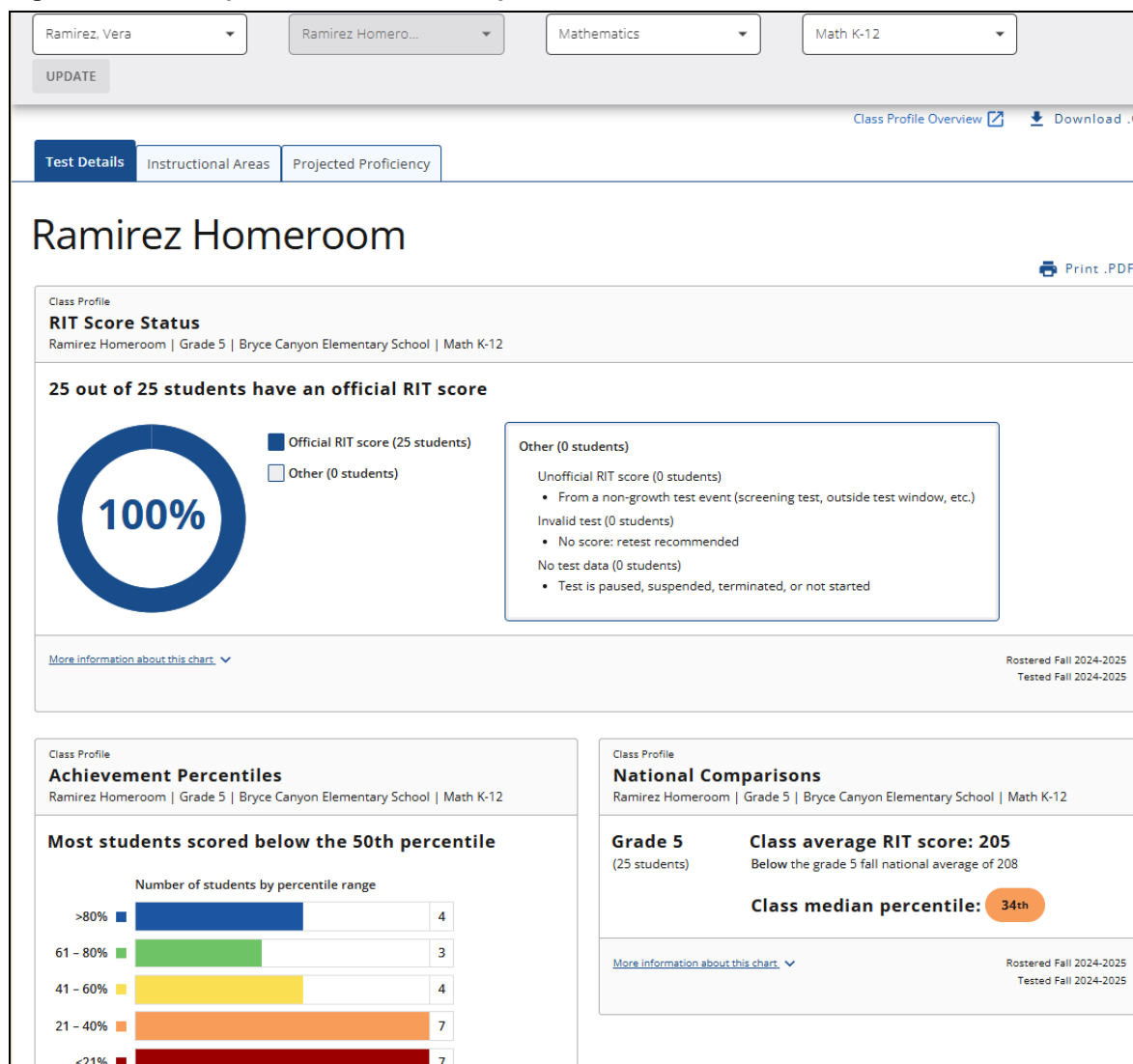
Class-level reports provide an overview of performance and detailed information about students within a class. Educators can use these reports to differentiate instruction for individual students or groups of students, inform classroom practice, and identify instructional areas of strength and opportunity for the class as a whole.

10.3.1. Class Profile Report

The Class Profile report is an interactive, actionable report that provides class- and student-level insights to support instructional planning. Features such as dynamic sorting and direct access to the Student Profile report allow educators to review MAP Growth results in easy-to-read views and use the information to guide instructional decisions.

As shown in the example presented in Figure 10.6, the Class Profile report helps educators quickly identify patterns in class performance and determine next steps for the class, group of students, or individual students.

Figure 10.6. Example of Class Profile Report



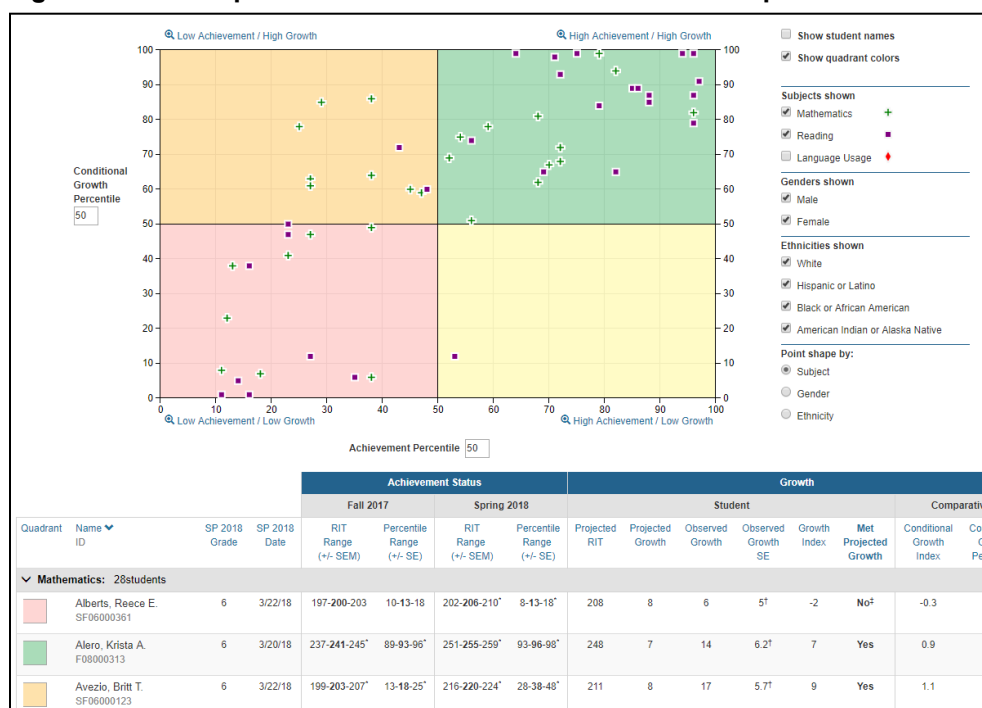
Data visualizations provide a high-level overview of class performance within the instructional areas on the selected MAP Growth test and help educators see how students' instructional area scores are distributed across achievement quintiles and 10-point RIT bands. This information supports a deeper understanding of class-level strengths and opportunities and provides a starting point for deciding which areas of the subject to emphasize during instruction. A brief video about using the Class Profile report is provided at <https://vimeo.com/537040943>.

10.3.2. Achievement Status and Growth Report

The Achievement Status and Growth report is useful for examining student learning, evaluating program effectiveness, and grouping students based on percentile information. This customizable report provides both static and interactive summaries of achievement and growth data. The static report shows each student's projected growth based on national norms and compares actual growth with projected growth.

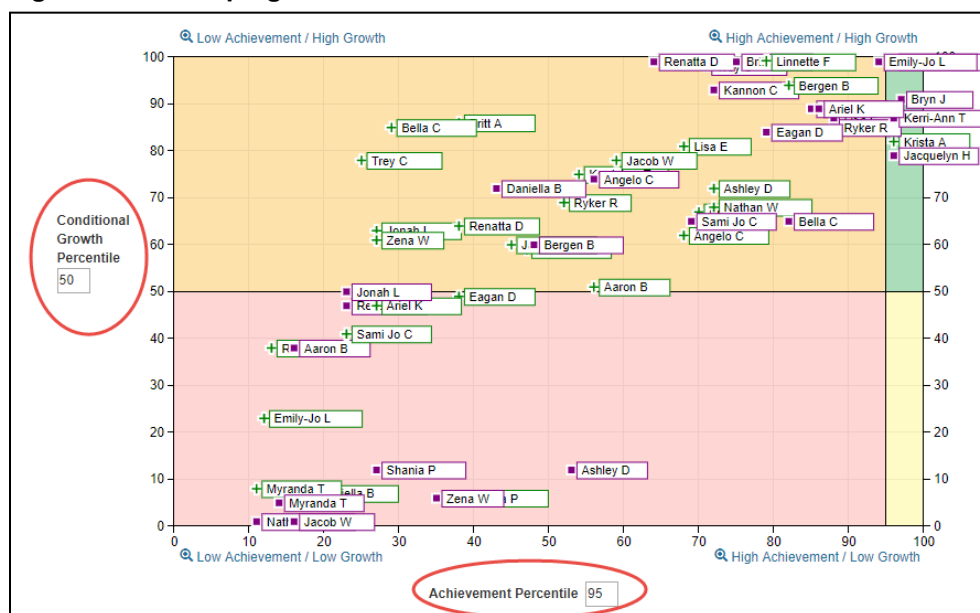
The interactive visualization allows educators to quickly review each student's achievement and growth, helping them target instruction more effectively. By default, achievement and growth are interpreted relative to the 50th percentile, as shown in Figure 10.7.

Figure 10.7. Example of Achievement Status and Growth Report



Using this report, educators can adjust the benchmarks used to compare achievement and growth, as demonstrated in Figure 10.8, to group students for more effective instruction, such as intervention or extension, or for other purposes, such as identifying students for a gifted program.

Figure 10.8. Grouping Students Based on Performance in Achievement Status and Growth Report



For additional information and a brief introductory video, please visit

https://teach.mapnwea.org/impl/maphelp/Content/Data/SampleReports/AchievementStatus_Growth.htm.

10.4. School and District Level Reports

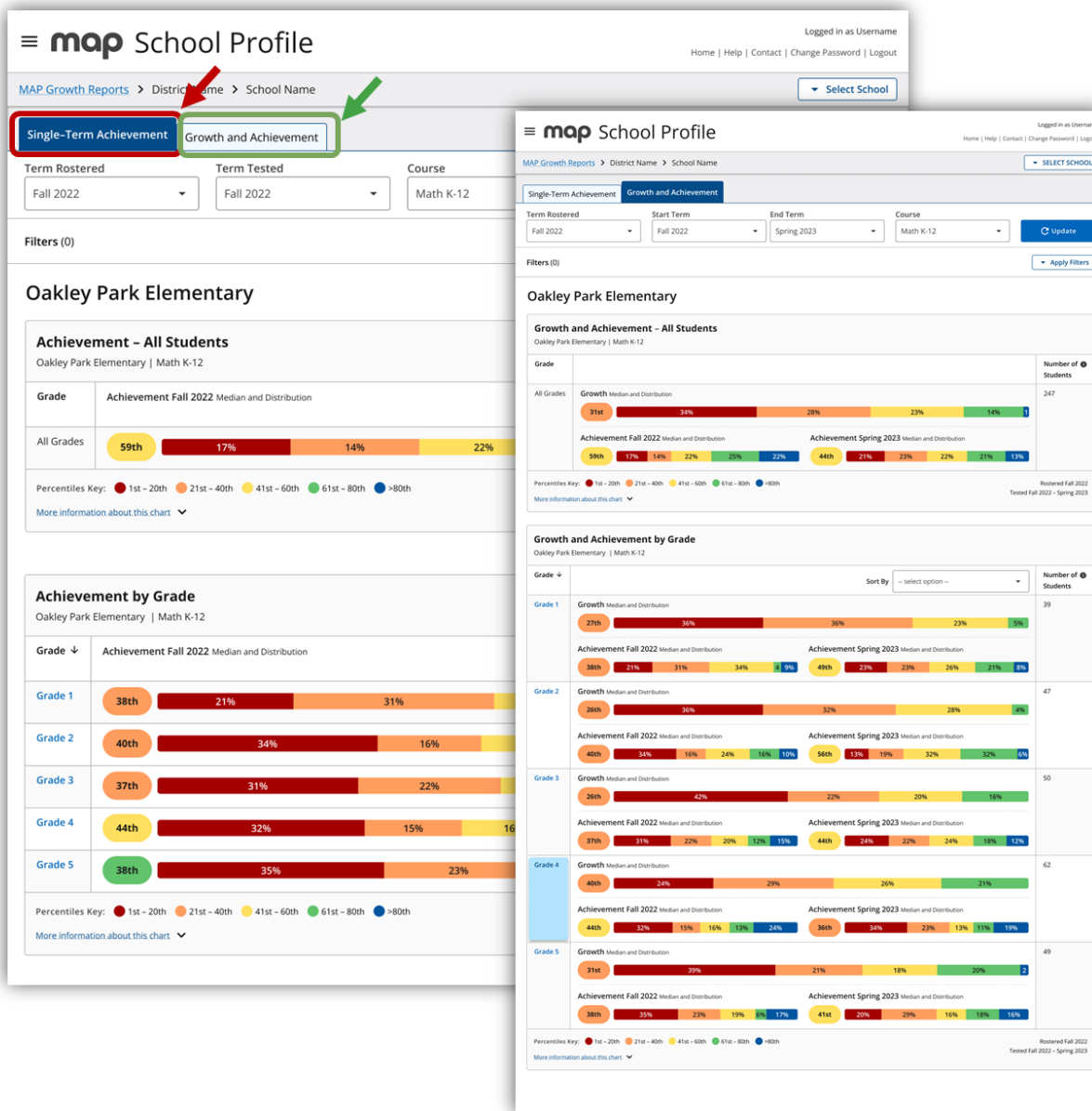
To help school administrators review trends, identify areas of strength and opportunity, and monitor the percentage of students meeting targets, NWEA provides aggregated reports at the school and district levels.

10.4.1. School Profile Report

The School Profile report provides school and district leaders with refinable data views for examining MAP Growth results at the school level. The report presents term-based growth medians and aggregate data by school and grade, helping leaders understand baseline performance and monitor changes in student outcomes over time.

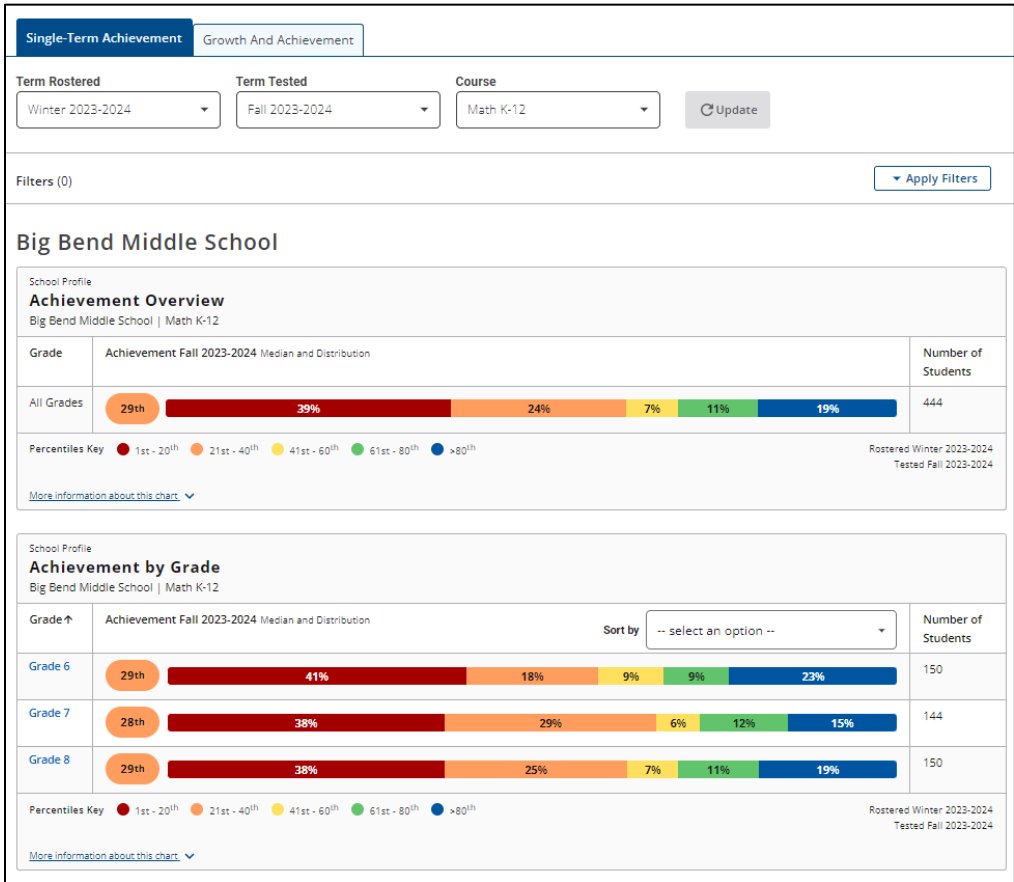
These insights can help leaders evaluate instructional practices and programs and identify areas where additional support may be needed. Data can be filtered by characteristics such as ethnicity, gender, and program, allowing school and district leaders to examine patterns across student groups. In addition to grade-level summaries, users can drill down to view class-level data within each grade. Figure 10.9 presents an example of a School Profile report.

Figure 10.9. Example School Profile Report



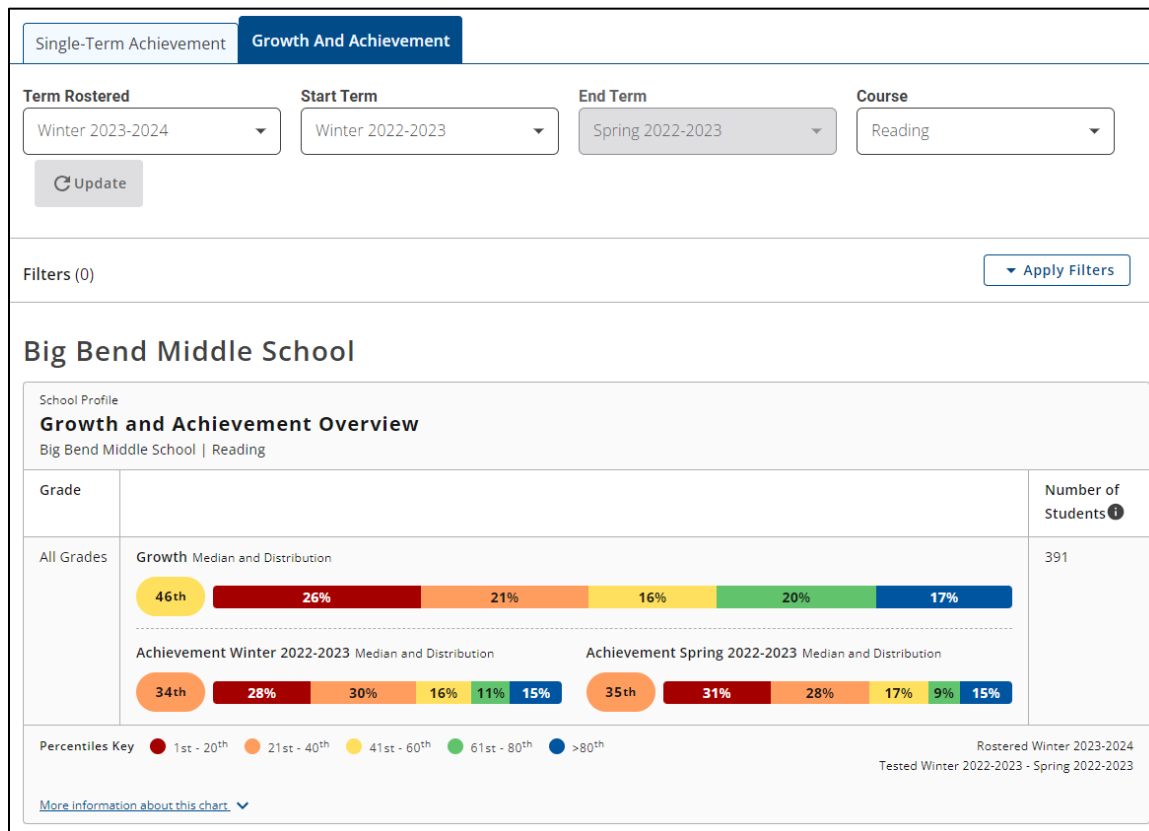
The Single-Term Achievement tab of the report, highlighted in Figure 10.10, presents achievement data for all grades in a school for a selected term. Educators can review the median achievement percentile, the distribution of students across achievement quintiles, and the number of students with valid growth measures for the selected population.

Figure 10.10. Example of Single-Term Achievement Tab in School Profile Report



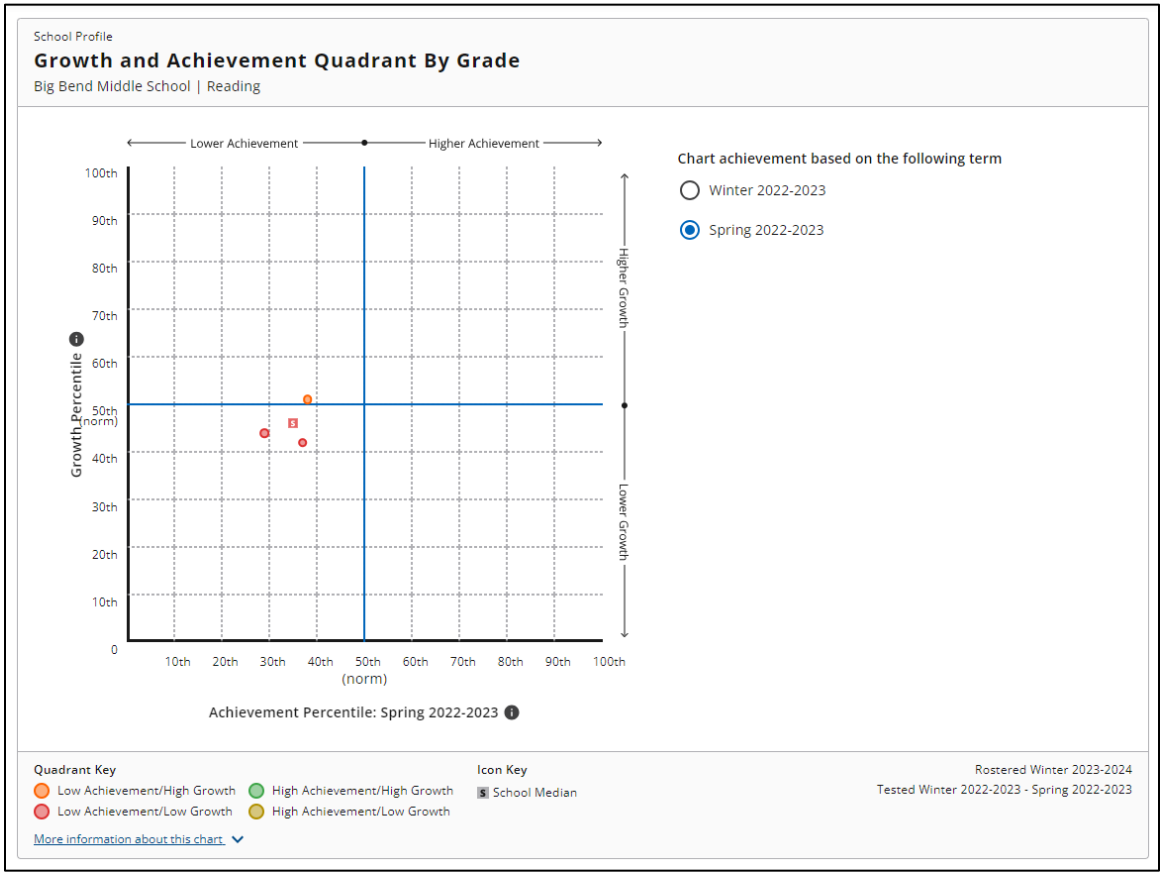
The Growth and Achievement tab of the School Profile report, highlighted in Figure 10.11, includes three modules that help educators explore data across two testing terms. The Growth and Achievement Overview module presents overall growth and achievement comparisons.

Figure 10.11. Example of Growth and Achievement Tab in School Profile Report



The Growth and Achievement Quadrant by Grade module, shown in Figure 10.12, displays growth and achievement data across two testing terms. The x-axis represents the grade's median achievement percentile for a selected term, and the y-axis represents the grade's median growth percentile between the start and end terms.

Figure 10.12. Example of Growth and Achievement Quadrant by Grade in School Profile Report



The Growth and Achievement by Grade module, shown in Figure 10.13, presents growth and achievement data by grade. Educators can select a grade to view class-level information or select a percentile band to review student-level details.

Figure 10.13. Example of Growth and Achievement by Grade in School Profile Report



10.4.2. District Profile Report

The District Profile report is designed to help district administrators examine performance trends by grade and school. The report supports monitoring of student achievement and growth over time, helping district leaders make decisions about programs, interventions, instructional supports, and curricular resources.

The District Profile report provides multiple levels of aggregated achievement and growth data, including districtwide summaries across all students, grades, and schools, as well as grade-level summaries across schools within the district.

Like the School Profile report, the District Profile report includes tabs for Single-Term Achievement, shown in Figure 10.14, and Growth and Achievement, shown in Figure 10.15. Data can be filtered by characteristics such as ethnicity, gender, and program. Educators can also select a grade to compare schools within that grade level.

Figure 10.14. Example of Single-Term Achievement Tab in District Profile Report

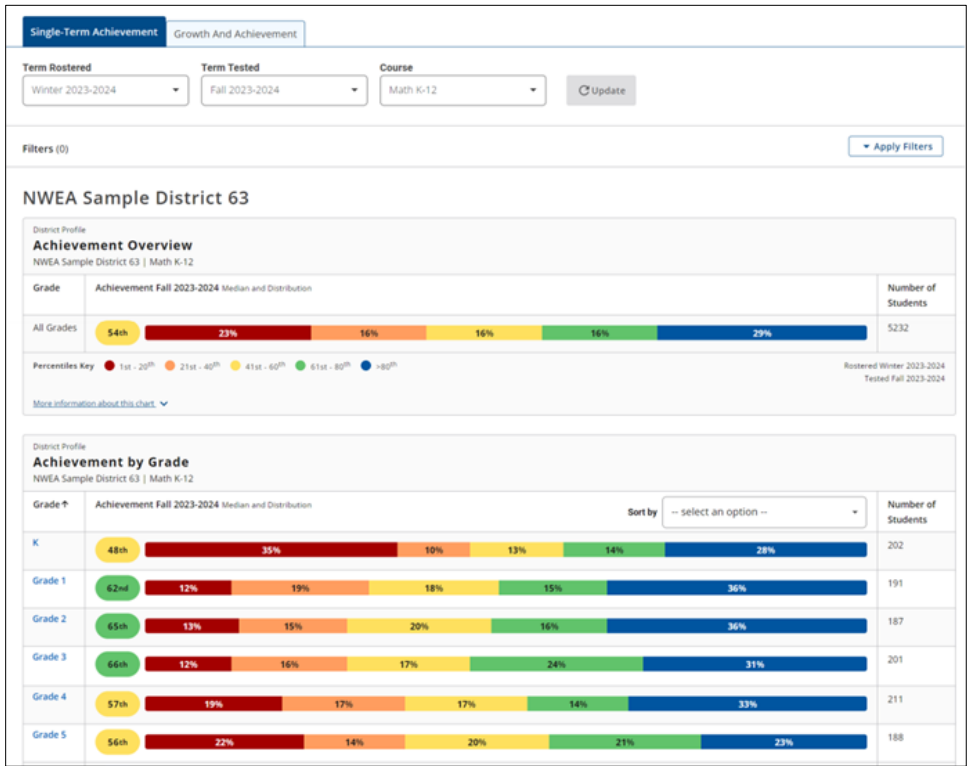
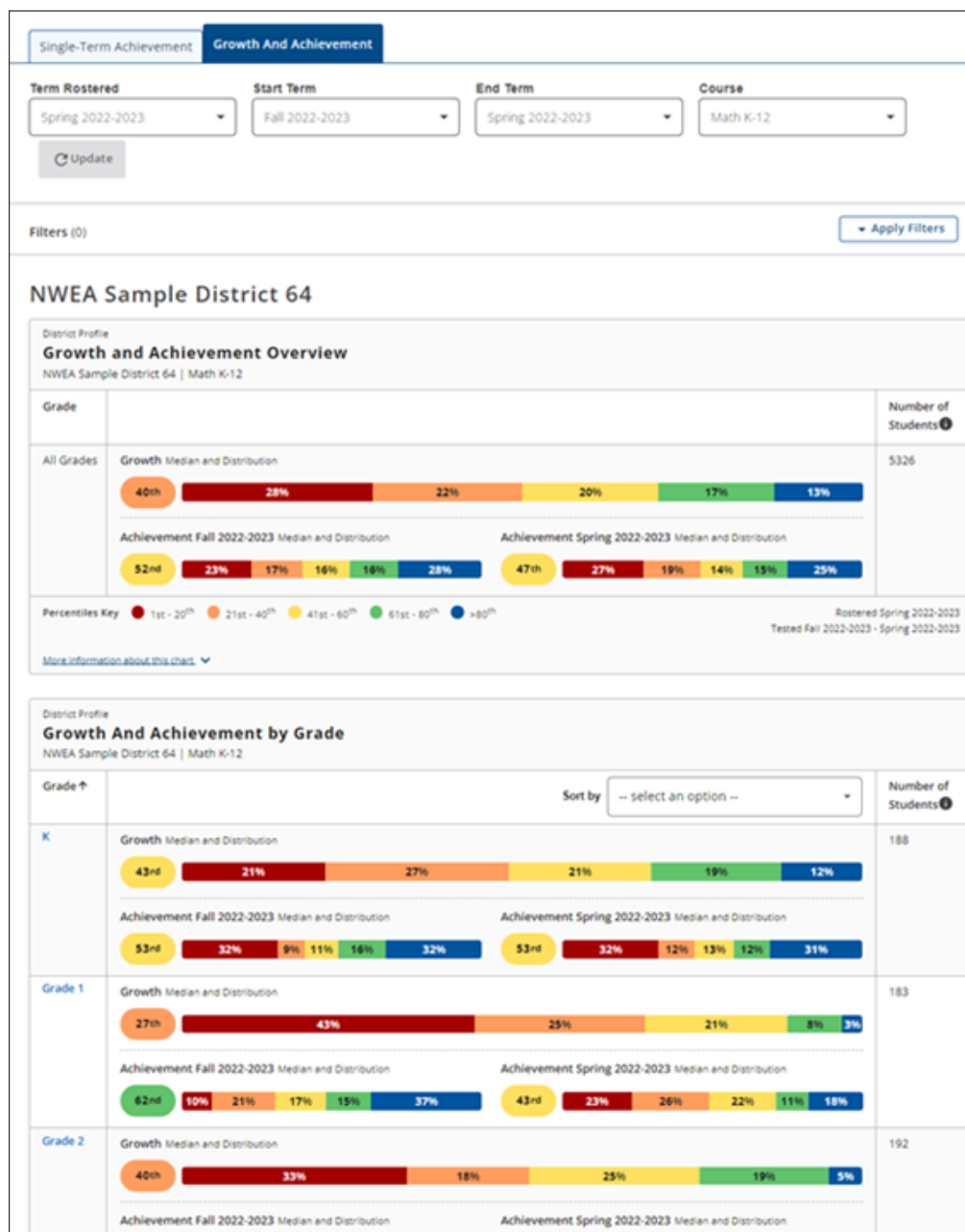


Figure 10.15. Example of Growth and Achievement Tab in District Profile Report



10.5. Score Interpretation Guide

The *MAP Growth Reports Portfolio*, NWEA's score interpretation guide for MAP Growth, provides descriptions and samples of available reports. Districts and schools can access the portfolio online at <https://www.nwea.org/resource-center/resource/map-growth-reports-portfolio>.

10.6. Professional Learning

NWEA professional learning helps educators and administrators build confidence in accessing, interpreting, using, and sharing assessment data. Through in-person, virtual, and asynchronous learning options, NWEA provides training and resources that support instructional decision-making and improved learning outcomes for students.

NWEA professional learning is evidence-based, grounded in the work of leading researchers, and aligned with the Learning Forward Standards for Professional Learning, the Interstate Teacher Assessment and Support Consortium Model Core Teaching Standards, and the Every Student Succeeds Act. Workshop participants consistently report high levels of satisfaction, with customer experience ratings of 93% or higher.

Table 10.1 provides a sample professional learning plan to support initial test administration and data analysis. Because each school district or educational agency has unique needs, NWEA works with school leaders to build professional learning plans that support long-term success. The workshops described in the table are part of the MAP Growth Learning Pathway, a series designed to help educators use MAP Growth data effectively to support student learning.

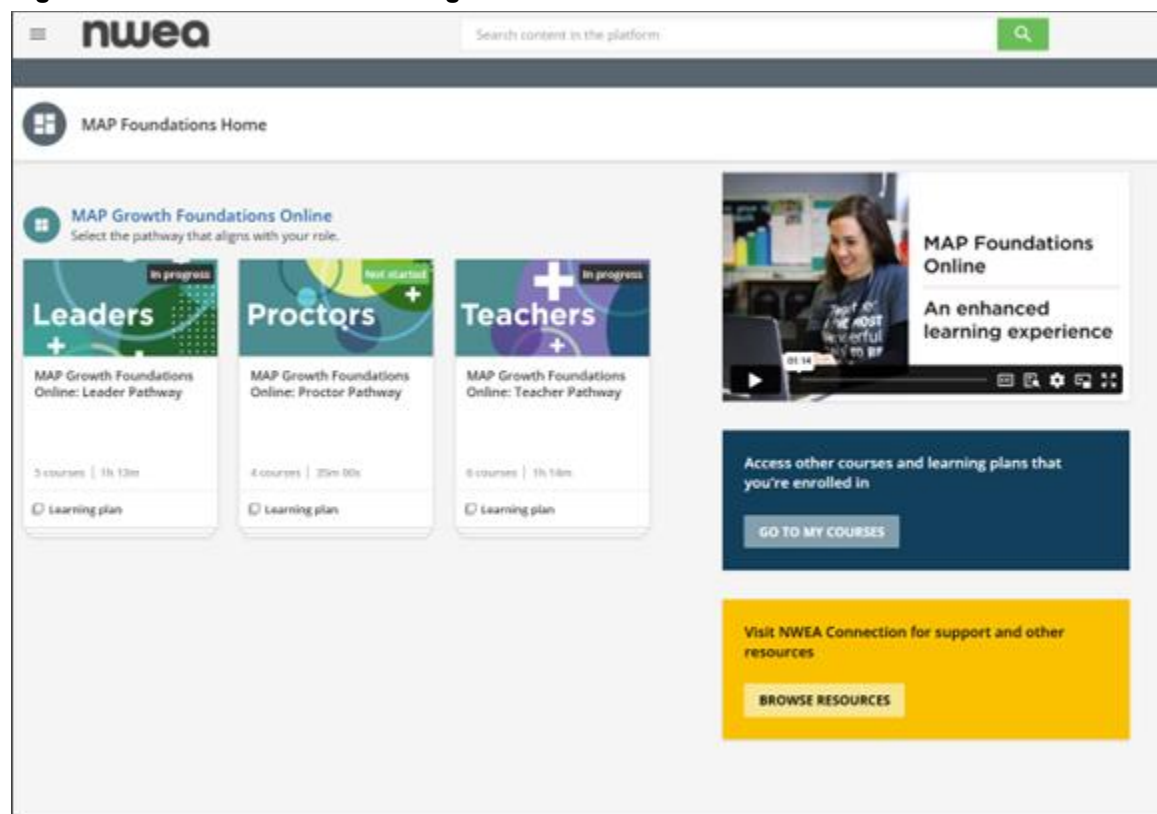
Table 10.1. Professional Learning Workshops

Workshop	Description
MAP Growth Foundations Online	Whether you're implementing MAP Growth for the first time, onboarding new staff, or simply refreshing your knowledge, Foundations Online provides the knowledge and tools staff across your school or district need to administer assessments, analyze reports, and act on data. This asynchronous online course provides learning pathways for teachers, administrators, and proctors. https://www.nwea.org/resource-center/brochure/74695/MAP-Growth-Foundations-Online_NWEA_brochure.pdf/
MAP Growth Basics	Gain a solid understanding of what makes MAP Growth unique, how to administer the assessments, and the importance of engaging students and leveraging data to inform instruction.
Applying Reports	Learn to access, interpret, and apply data from MAP Growth reports and how to use the data to inform ongoing work, with a particular focus on goal setting with students.
Informing Instruction	Support effective instruction and meet the needs of every student through responsive instruction using MAP Growth results.
Focusing on Growth	Learn how to apply MAP Growth data in goal setting and data conversations to improve student learning.
Action Planning with MAP Data	Learn how to use MAP Growth data to supercharge your instruction and school-improvement goals. Educators will get tools and strategies for turning data from MAP Growth into meaningful action.
Data Coaching and Consulting	A dedicated coach can help you build local capacity, establish constructive and ongoing data conversations, expand assessment literacy district-wide, and more.

10.7. On-Demand Learning

In addition to the blended learning approach described in Table 10.1, on-demand training is available to all staff through Professional Learning Online (see Figure 10.16). This one-stop eLearning site empowers educators to access training and resources at their pace and on their schedule. It incorporates a wide range of activities, from learning the basics of NWEA's assessments to using data to support student learning. Professional Learning Online is available from any location with an internet connection. See also https://www.nwea.org/resource-center/brochure/55984/Professional-learning_NWEA_brochure.pdf/

Figure 10.16. Professional Learning Online



10.8. Additional Resources for Developing Understanding

10.8.1. Online Help Center

MAP Growth assessments include online help resources and troubleshooting support embedded in the system. These materials include step-by-step training videos and guides for proctors, educators, and administrators.

10.8.2. NWEA Connection

[NWEA Connection](#) is an online support resource center where educators and education leaders can find information, ask questions, and access support for NWEA products and services, including live chat support. NWEA Connection is part of a comprehensive approach to customer support and professional learning.

On NWEA Connection, educators can:

- Chat with an NWEA Partner Support representative
- Access content that supports successful assessment use
- Search the Support Knowledge Base for answers to product questions
- Review updates and changes to NWEA products
- View support ticket status
- Securely connect with NWEA experts and partners on specific topics

10.8.3. Resource Center

The [Resource Center](#) at NWEA.org provides a variety of resources for educators and administrators on a wide range of topics. Available materials include webinars, eBooks, white papers, case studies, guides, briefs, and other resources. The site also includes the *Teach. Learn. Grow.* blog, which offers tips and discussions on topics relevant to educators.

10.8.4. NWEA YouTube Channel

The NWEA YouTube channel provides short videos, recorded webinars, deep-dive resources, and curated playlists on topics such as MAP Growth 101 and Norms 101. These resources offer another way for educators and administrators to learn about MAP Growth and related NWEA products. Additional resources are available on the NWEA YouTube channel: <https://www.youtube.com/@NWEAvideos>.

References

- Achieve, Inc. (2019). *A framework to evaluate cognitive complexity in mathematics assessments*.
https://www.achieve.org/files/Mathematics%20Cognitive%20Complexity%20Framework_Final_92619.pdf
- 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
- Anderson, L. W., & Krathwohl, D. R. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy for educational objectives*. Longman.
- Atteberry, A., & McEachin, A. (2021). School's out: The role of summers in understanding achievement disparities. *American Educational Research Journal*, 58(2), 239–282.
<https://doi.org/10.3102/0002831220937285>
- Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34(3), 197–211. <https://doi.org/10.1111/j.1745-3984.1997.tb00515.x>
- Center for Applied Special Technology (CAST) (2018). *Universal design for learning guidelines version 2.2* [Infographic]. CAST.
https://udlguidelines.cast.org/static/udlg_graphicorganizer_v2-2_numbers-yes.pdf
- Council of Chief State School Officers (CCSSO). (2016, August). *CCSSO accessibility manual: How to select, administer, and evaluate use of accessibility supports for instruction and assessment of all students*. CCSSO.
- Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62(2), 369–383. <https://doi.org/10.1348/000711008X304376>
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum Associates, Inc. <https://doi.org/10.1002/j.2333-8504.1992.tb01440.x>
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, 23(4), 355–368.
<https://www.jstor.org/stable/1434554>
- Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement*, 40(4), 289–301. <https://doi.org/10.1177/0146621616631317>

- He, W. (2022). *MAP Growth item parameter drift study*. NWEA Research Report. NWEA. <https://www.nwea.org/uploads/2021/06/MAP-Growth-Item-Parameter-Drift-2022-01-14.pdf>
- He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*, 74(4), 677–696. <https://doi.org/10.1177/0013164413517503>
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc.
- Hu, A. (2021). *MAP Growth linking studies: Intended uses, methodology, and recent studies*. NWEA. https://www.nwea.org/uploads/2021/06/MAP-Growth-Linking-Studies_Uses-Methodology-Recent-Studies-2021-11-15.pdf
- Ingebo, G. S. (1997). *Probability in the measure of achievement*. MESA Press.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedure for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6
- Linacre, J. M. (2023). *Winsteps*® (Version 5.6.0) [Computer software]. Winsteps.com. Available from <https://www.winsteps.com/>
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748. <https://doi.org/10.1093/jnci/22.4.719>
- McNeish, D., & Dumas, D. (2021). A seasonal dynamic measurement model for summer learning loss. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 184(2), 616–642. <https://doi.org/10.1111/rssa.12634>
- Nordengren, C. (2023). *Focusing squarely on students: A theory of change for NWEA learning and improvement services*. NWEA. <https://www.nwea.org/resource-center/resource/focusing-squarely-on-students-a-theory-of-change-for-nwea-professional-learning/>
- NWEA (2022). *Achievement and growth norms for course-specific MAP Growth tests*. NWEA. <https://www.nwea.org/uploads/2023/02/2022-MAP-Growth-HS-Math-Norms-Report-02-07-2023.pdf>
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70(350), 351–356. <https://doi.org/10.2307/2285821>
- Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2015, November). *Continued progress: Promising evidence on personalized learning*. RAND Corporation. <https://www.jstor.org/stable/10.7249/j.ctt19w72z1>

- Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5–13. <https://doi.org/10.1111/j.1745-3992.2009.00149.x>
- Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, 14(1), 58–93. <https://doi.org/10.1163/24689300-01401006>
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press.
- Shin, C. D., Chien, Y., Way, W. D., & Swanson, L. (2009, April). *Weighted penalty model for content balancing in CATS*. Pearson.
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17(3), 277–292. <https://psycnet.apa.org/doi/10.1177/014662169301700308>
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes. <https://nceo.umn.edu/docs/onlinepubs/synth44.pdf>
- von Hippel, P. T., & Hamrock, C. (2019). Do test score gaps grow before, during, or between the school years? Measurement artifacts and what we can know in spite of them. *Sociological Science*, 6(3), 43–80. <http://dx.doi.org/10.15195/v6.a3>
- Webb, N. L. (1997). Criteria for alignment of expectations and assessments on mathematics and science education. *Research Monograph No. 6*. Council of Chief State School Officers and National Institute for Science Education (NISE). University of Wisconsin-Madison. <https://mspnet-static.s3.amazonaws.com/WebbCriteria.pdf>
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.
- Yoon, A. J., & Merry, J. J. (2017). Understanding the role of schools in the Asian-white gap: A seasonal comparison approach. *Race Ethnicity and Education*, 21(5), 680–700. <https://doi.org/10.1080/13613324.2017.1365053>
- Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates, Inc.
- Zwick, R., & Mislevy, R. J. (2011, March). *Scaling and linking through-course summative assessments*. [Paper presentation]. Invitational Research Symposium on Through-Course Summative Assessment, Atlanta, GA, United States. https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Zwick_Mislevy.pdf

Appendix A: Marginal Reliability by Course, State, and Term

Table A1. Marginal Reliability by Course, State, and Term

Course	State	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Algebra 1	AL	1,790	0.96	1,443	0.98	1,417	0.98
	CA	3,634	0.95	3,033	0.96	2,970	0.96
	CO	1,305	0.96	1,106	0.96	1,179	0.97
	DC	1,929	0.96	1,892	0.97	1,751	0.98
	FL	7,130	0.95	5,531	0.95	3,295	0.96
	IL	2,407	0.96	2,269	0.95	2,233	0.96
	IN	4,213	0.95	3,195	0.95	3,261	0.96
	MD	1,311	0.96	1,231	0.97	1,056	0.98
	MI	3,864	0.96	3,117	0.96	2,830	0.97
	MO	10,176	0.95	9,922	0.96	6,838	0.96
	NJ	4,352	0.97	2,323	0.97	4,332	0.98
	NV	1,581	0.95	1,399	0.96	1,363	0.97
	NY	31,887	0.97	26,464	0.98	23,418	0.98
	OH	20,280	0.96	18,158	0.97	10,512	0.97
	PA	5,074	0.97	4,245	0.98	3,721	0.98
	SC	1,943	0.97	1,263	0.98	1,390	0.97
	TX	85,655	0.96	70,566	0.97	77,690	0.97
	VA	11,732	0.96	7,868	0.96	6,483	0.97
Algebra 2	CA	4,029	0.96	3,804	0.97	3,418	0.97
	DC	2,136	0.97	2,325	0.98	2,134	0.98
	FL	2,452	0.96	1,706	0.97	2,180	0.97
	IN	1,855	0.95	1,640	0.95	1,589	0.95
	MI	2,997	0.95	2,285	0.96	1,611	0.96
	MO	1,947	0.95	1,678	0.96	1,125	0.97
	NJ	4,743	0.97	2,026	0.97	4,203	0.98
	NY	20,496	0.98	17,718	0.98	15,376	0.98
	OH	5,596	0.96	4,825	0.97	3,365	0.97
	TX	61,787	0.96	47,141	0.97	56,596	0.97
	VA	12,724	0.97	6,306	0.97	5,610	0.98
Geometry	CA	3,862	0.95	3,630	0.96	3,689	0.97
	CO	1,244	0.96	1,094	0.97	1,023	0.97
	DC	2,641	0.96	2,568	0.97	2,564	0.98
	FL	9,443	0.95	8,668	0.96	4,643	0.97
	GA	2,627	0.97	1,464	0.97	1,951	0.98
	IL	1,811	0.95	1,632	0.96	1,739	0.97
	IN	2,696	0.94	2,490	0.95	2,613	0.96
	MI	3,987	0.95	1,584	0.96	3,339	0.96
	MO	3,177	0.95	2,577	0.96	2,181	0.97

Course	State	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
	NJ	4,606	0.96	2,196	0.96	4,470	0.98
	NY	22,726	0.96	20,284	0.97	16,125	0.98
	OH	17,339	0.96	16,248	0.97	9,020	0.97
	TX	67,630	0.96	53,996	0.97	63,016	0.98
	VA	21,882	0.96	12,266	0.96	10,076	0.97
Integrated Math 1	CA	8,693	0.96	4,820	0.97	8,091	0.97
Integrated Math 2	CA	8,105	0.96	4,321	0.96	7,292	0.97
Integrated Math 3	CA	2,751	0.96	2,130	0.97	2,261	0.98
Biology/ Life Science	DC	1,413	0.96	1,458	0.96	1,410	0.97
	MO	4,360	0.94	3,981	0.95	2,519	0.96
	OH	8,876	0.94	8,420	0.95	5,457	0.96
	PA	1,668	0.95	1,552	0.95	1,456	0.96
	TX	86,127	0.94	69,263	0.95	78,358	0.96

Note. Results for a state were omitted if the sample size for any term was less than 1,000.

Appendix B: Marginal Reliability by Course, Instructional Area, and Term

Table B1. Marginal Reliability by Course, Instructional Area, and Term

Course	Instructional Area	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Algebra 1	Algebra	43,722	0.90	37,579	0.91	31,407	0.92
	Algebraic Reasoning	6,424	0.83	5,499	0.87	3,205	0.88
	Computation and Estimation with Real Numbers	9,054	0.86	6,758	0.88	6,989	0.90
	Data Analysis & Statistics	4,181	0.76	3,165	0.78	3,261	0.79
	Data Analysis and Probability	6,424	0.80	5,499	0.84	3,205	0.84
	Describe & Graph Linear Functions, Equations, & Inequalities	85,031	0.85	70,441	0.87	77,601	0.89
	Descriptive Statistics	37,371	0.86	27,743	0.86	26,151	0.87
	Equations and Inequalities	57,104	0.85	40,493	0.87	38,137	0.89
	Expressions	9,054	0.86	6,758	0.88	6,989	0.91
	Expressions and Operations	9,597	0.88	5,467	0.89	4,747	0.91
	Functions	106,852	0.87	88,040	0.89	71,376	0.91
	Functions and Statistics	9,597	0.89	5,467	0.91	4,747	0.92
	Linear Equations, Inequalities, and Functions	4,181	0.87	3,165	0.90	3,261	0.90
	Number & Data	10,028	0.85	9,866	0.87	6,717	0.90
	Number & Quantity	31,784	0.89	26,450	0.90	23,332	0.91
	Number Sense, Operations, and Financial Literacy	6,424	0.62	5,499	0.70	3,205	0.76
	Number Sense, Properties, and Number Theory	9,054	0.86	6,758	0.88	6,989	0.90
	Number Systems, Expressions, and Functions	4,181	0.80	3,165	0.83	3,261	0.88
	Number and Algebraic Methods	85,031	0.87	70,441	0.89	77,601	0.92
	Number and Quantity; Interpreting Data	1,910	0.90	1,263	0.94	1,359	0.93
	Number, Quantities, Equations, and Expressions	18,200	0.90	16,538	0.92	10,322	0.92

Course	Instructional Area	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
	Numerical and Algebraic Expressions	37,979	0.85	28,180	0.87	26,345	0.89
	Patterns, Functions, and Graphing	9,054	0.87	6,758	0.89	6,989	0.92
	Quadratic and Exponential Functions and Equations	85,031	0.66	70,441	0.72	77,601	0.81
	Statistics	18,200	0.89	16,538	0.90	10,322	0.90
	Statistics & Probability	31,784	0.89	26,450	0.90	23,332	0.91
	Systems; Quadratic and Exponential Equations and Functions	4,181	0.71	3,165	0.66	3,261	0.76
	Write & Solve Linear Functions, Equations, & Inequalities	85,031	0.88	70,441	0.90	77,601	0.91
Algebra 2	Algebra	1,947	0.83	1,677	0.87	1,125	0.90
	Algebraic Reasoning	2,301	0.89	1,650	0.92	2,000	0.93
	Computation and Estimation with Real Numbers	3,539	0.86	2,738	0.88	2,968	0.89
	Descriptive Statistics	31,216	0.88	24,500	0.89	22,717	0.91
	Equations and Inequalities	47,735	0.88	33,595	0.89	31,427	0.90
	Expressions	3,539	0.86	2,738	0.88	2,969	0.88
	Expressions and Operations	12,734	0.90	6,308	0.91	5,614	0.92
	Function Attributes, Inverses, and Data	61,511	0.87	47,050	0.90	56,463	0.91
	Functions	56,335	0.89	45,616	0.91	41,365	0.92
	Functions and Statistics	12,722	0.91	6,307	0.92	5,610	0.94
	Number & Data	1,947	0.88	1,676	0.92	1,125	0.92
	Number & Quantity and Algebra	20,489	0.92	17,715	0.94	15,356	0.94
	Number Sense, Operations, and Financial Literacy	2,301	0.87	1,650	0.91	2,000	0.90
	Number Sense, Properties, and Number Theory	3,539	0.85	2,738	0.87	2,969	0.87
	Number and Algebraic Methods	61,511	0.84	47,050	0.87	56,463	0.89

Course	Instructional Area	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
	Numerical and Algebraic Expressions	31,519	0.81	24,552	0.85	22,875	0.86
	Other Functions, Equations, Inequalities, and Systems	61,511	0.87	47,050	0.89	56,463	0.90
	Patterns, Functions, and Graphing	3,539	0.85	2,738	0.87	2,970	0.88
	Quadratic, Square Root, Exponential, and Logarithmic Functions, Equations, and Inequalities	61,511	0.83	47,050	0.88	56,463	0.89
	Statistics & Probability	20,489	0.92	17,715	0.93	15,356	0.94
Geometry	Applications of Probability	29,251	0.86	20,671	0.88	24,312	0.90
	Circles	17,298	0.86	16,243	0.87	8,976	0.90
	Circles and Geometric Properties with Equations	22,643	0.83	20,263	0.86	16,063	0.90
	Congruence	22,643	0.87	20,263	0.90	16,062	0.91
	Congruence and Proof	17,298	0.85	16,243	0.87	8,975	0.88
	Congruence, Similarity, Right Triangles, & Trig	29,650	0.82	20,836	0.86	24,505	0.89
	Coordinate Geometry and Circles	9,238	0.85	8,609	0.88	4,444	0.90
	Coordinate and Transformational Geometry	67,335	0.86	53,798	0.87	62,833	0.88
	Data & Statistical Reasoning; Probabilistic Reasoning	1,710	0.90	1,343	0.91	1,246	0.91
	Geometric & Spatial Reasoning	1,710	0.89	1,343	0.90	1,246	0.91
	Geometric Measurement and Modeling	52,235	0.86	41,050	0.88	40,528	0.90
	Geometric Measurement and Theorems	9,238	0.86	8,609	0.89	4,444	0.91
	Geometric Properties with	29,647	0.76	20,830	0.79	24,506	0.84

Course	Instructional Area	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
	Equations and Circles						
	Geometric Relationships	8,082	0.89	6,337	0.91	6,401	0.92
	Geometry Foundations	2,695	0.81	2,490	0.86	2,610	0.88
	Logic, Circles, and Probability	67,335	0.88	53,792	0.90	62,818	0.93
	Measurement of Geometric Shapes	8,082	0.90	6,337	0.91	6,388	0.92
	Patterning & Algebraic Reasoning	1,710	0.90	1,343	0.90	1,246	0.90
	Polygons, Circles, and Three-Dimensional Figures	21,814	0.85	12,278	0.86	10,023	0.91
	Probability	17,298	0.86	16,243	0.88	8,976	0.90
	Proof, Congruence, Similarity, and Trigonometry	67,335	0.85	53,798	0.88	62,832	0.90
	Properties of Geometric Shapes	8,082	0.88	6,337	0.90	6,394	0.91
	Properties of Transformations and Trigonometry	9,238	0.85	8,609	0.86	4,444	0.90
	Reasoning, Lines and Transformations	21,814	0.90	12,278	0.90	10,023	0.91
	Similarity & Trigonometry	17,298	0.84	16,243	0.87	8,976	0.90
	Similarity, Right Triangles, and Trigonometry	22,643	0.86	20,262	0.89	16,061	0.92
	Transformations & Three-Dimensional Solids	2,695	0.81	2,490	0.79	2,610	0.83
	Triangles	24,509	0.88	14,768	0.88	12,633	0.90
	Two-Dimensional Shapes	2,695	0.78	2,490	0.82	2,610	0.86
	Two-Dimensional and Three-Dimensional Figures	67,334	0.84	53,797	0.85	62,830	0.89
Integrated Math 1	Algebra and Quantities	12,169	0.82	6,761	0.88	9,620	0.89
	Descriptive Statistics	11,770	0.86	6,704	0.88	9,604	0.88
	Functions	12,110	0.83	6,761	0.89	9,616	0.90
	Geometry	13,339	0.83	8,132	0.86	10,463	0.86

Course	Instructional Area	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Integrated Math 2	Algebra and Number	10,086	0.82	5,952	0.86	8,159	0.87
	Applications of Probability	9,801	0.86	5,895	0.89	8,113	0.90
	Functions	10,021	0.81	5,934	0.85	8,157	0.87
	Geometry	10,102	0.81	6,074	0.86	8,481	0.88
Integrated Math 3	Algebra and Number	4,746	0.86	3,758	0.89	3,077	0.90
	Descriptive Statistics	4,671	0.88	3,741	0.89	3,050	0.91
	Functions	4,701	0.85	3,750	0.88	3,063	0.89
	Geometry	4,924	0.88	3,901	0.88	3,224	0.90
Biology/Life Science	Biological Structures, Functions, and Processes	86,059	0.82	69,255	0.86	78,293	0.89
	Cells; Heredity; Evolution	8,693	0.89	8,386	0.90	5,452	0.92
	Diversity and Interdependence of Life	8,693	0.90	8,386	0.90	5,452	0.91
	Ecosystems: Interactions, Energy, and Dynamics	20,292	0.83	16,031	0.87	12,066	0.88
	From Molecules to Organisms: Structures and Processes	20,072	0.82	15,984	0.86	12,010	0.87
	Heredity: Inheritance and Variation of Traits; Biological Evolution: Unity and Diversity	20,031	0.83	15,978	0.85	12,007	0.87
	Interdependence within Environmental Systems	86,059	0.83	69,255	0.85	78,293	0.87
	Mechanisms of Genetics, Biological Evolution	86,059	0.83	69,255	0.87	78,293	0.89

Note. Results for an instructional area were omitted if the sample size for any term was less than 1,000.