MAP Growth Technical Report for 2024–2025

April 2, 2026

nwea®

MAP Growth 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 is an interim computer adaptive test administered multiple times per year (fall, winter, spring, summer) in Math, Reading, Language Usage, and Science. Scores are reported on the RIT scale, an equal-interval vertical scale that allows comparison across grades and terms. The adaptive nature of MAP Growth enables efficient and precise measurement by prioritizing content and tailoring item difficulty to student ability, maximizing information and minimizing test length while maintaining measurement precision. The system supports both in-school and remote testing environments, with trained proctors facilitating test administration.

This document is a comprehensive technical report examining MAP Growth 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 the particular claim. The claims examined are listed below, along with a summary of the evidence provided in each chapter. Separate reports describe the Spanish MAP Growth offerings and the course-specific assessments.

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, and the NWEA Content Frameworks. The item pool for each grade-banded, standards-aligned test is created by selecting items aligned to a particular set of content standards and mapping them to cross-grade instructional areas, ensuring coverage beyond the tested grade band to accommodate students performing below or above grade level. 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, item sets, 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 Off-Grade in a Principled Way

MAP Growth employs a computer adaptive testing algorithm that selects items based on a student's estimated ability, providing more-difficult items following correct responses and easier items following incorrect responses. Content targeting and prioritization ensures that test blueprints are met by selecting items for each instructional area. The Enhanced Item Selection Algorithm prioritizes on-grade items but allows off-grade items to be selected as needed to match student ability, with low-performing students receiving more below-grade items and high-performing students receiving more above-grade items. Simulations and real data analyses confirm that the adaptive engine produces tests that satisfy blueprint requirements and provide precise measurement of student scores.

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

Score reliability is examined as the consistency of test scores for groups of students, expressed as the ratio of true score variance to observed score variance. MAP Growth reliability estimates include internal consistency, estimated by marginal reliability in IRT, and test-retest reliability, which accounts for the influence of time on scores. Across subjects and grades, reliability coefficients are generally high and average about 0.95, with slightly lower values in Science and early grades. Test-retest reliability is stronger for shorter intervals (winter–spring) than longer intervals (fall–spring). Measurement error is quantified by the conditional standard error of measurement (CSEM), which remains relatively constant across ability deciles, except for slightly higher values in the extreme deciles.

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. Norms allow comparison with a nationally representative sample, 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 *MAP Growth Technical Report* provides extensive evidence that MAP Growth assessments are psychometrically sound, fair, accessible, and useful tools for measuring student achievement and growth across diverse populations and educational settings.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. A Brief History of MAP Growth

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

The company was originally named the Northwest Evaluation Association and was later renamed simply NWEA. HMH (formerly named Houghton Mifflin Harcourt) acquired NWEA in 2023. Since its founding, NWEA has developed assessments that enable educational agencies to measure student learning with a high degree of accuracy in a relatively short amount of time. NWEA is perhaps best known for its MAP Growth assessment, but it produces other assessments including MAP Reading Fluency and several state assessment programs that combine summative and interim assessments.

Measures of Academic Progress (MAP) was officially introduced by NWEA in 2000 as the first large-scale, standardized computerized adaptive tests 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. Tests in each subject area are vertically scaled, allowing scores to be compared across terms and grade levels. In 2006, NWEA expanded its offerings in Reading and Mathematics by launching Measures of Academic Progress for Primary Grades (MAP for Primary Grades). It was designed to be a developmentally appropriate assessment for students in kindergarten through grade 2.

Early implementations of MAP required local installation of testing software on school computers, and assessments were administered over local school networks without the need for continuous internet connectivity. By the late 2000s, as internet infrastructure improved, MAP Growth transitioned to a fully web-based platform, enabling more streamlined administration and increased accessibility. In 2025, NWEA fully implemented another notable revision to MAP with its Enhanced Item Selection Algorithm (EISA).

Beginning in the late 2010s, both MAP and MAP for Primary Grades underwent multiple rebranding efforts. Measures of Academic Progress was rebranded as MAP Growth, and MAP for Primary Grades was renamed MAP Growth K–2. Collectively, MAP Growth utilizes age-appropriate content with grade-banded assessments with a common vertical scale in each subject.

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 spans multiple administrations across grades K–12 in Math and Reading and grades 2–12 in Language Usage and Science. MAP Growth is a computer adaptive test (CAT) that accounts for statistical and content requirements. Its adaptive engine selects items with a difficulty value matching a student's performance level while prioritizing grade-level content to achieve the goal of delivering test events that meet test blueprints. Through innovative test design and scoring methods, MAP Growth provides information about student achievement status and growth along with predictions of end-of-year proficiency. National comparisons at the student and school levels are provided through the 2025 MAP Growth norms.

The MAP Growth theory of action describes key features of MAP Growth and its position in a comprehensive assessment system. The basic premise to the theory of action is that when MAP Growth is situated in a comprehensive assessment system and used for its intended purposes (to yield information about student learning and enable educators to make data-informed decisions about curriculum and instruction), all students learn. Figure 1.1 illustrates the components of the MAP Growth theory of action, shedding light on the connectedness of the parts and the claims that are central to the validity argument. Evidence for the claims is described throughout this technical report and within other 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.

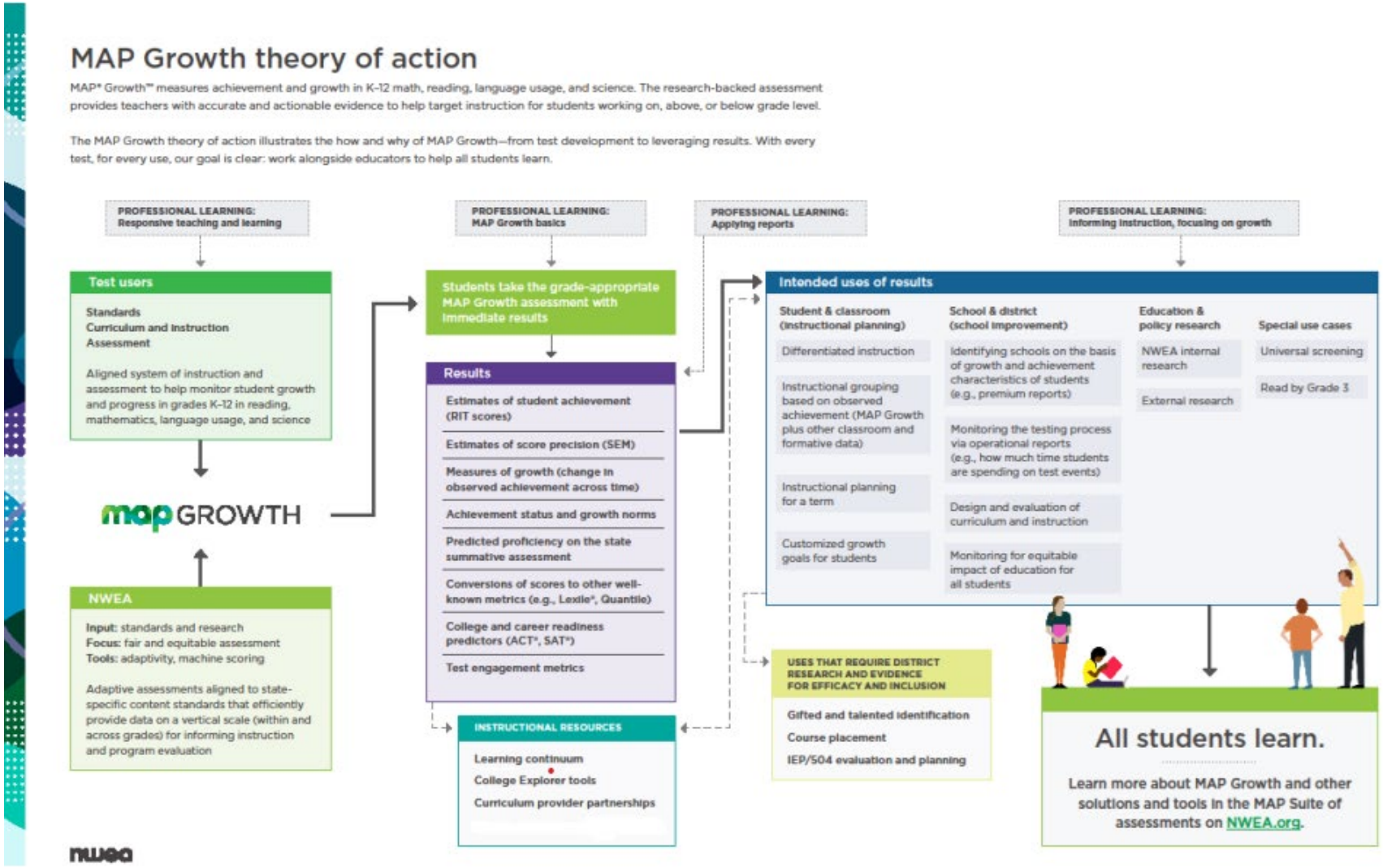
Although MAP Growth is not a summative assessment in and of itself, it has been tightly integrated with summative assessments in several states. In these innovative programs, MAP Growth tests are given in fall and winter. A short summative assessment is then given in the spring, supplemented with additional 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]

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.

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

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



1.2.2. MAP Growth Assessment Design

MAP Growth tests are aligned to content standards in a way that supports the cross-grade vertical scaling of the assessment. Instructional areas are established according to standards to reflect the articulation of content across grades. Tests are also grade banded (e.g., K–2, 2–5, 6+) to ensure that item format and test content are suitable for students from different grade levels. 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, common stimulus, 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 scores are reported on the Rasch Unit (**R**asch un**IT**) scale that ranges from about 100 to 350 with a mean of 200 and a standard deviation of 10. The RIT scale has two key characteristics. It is a vertical scale with equal-interval units. The vertical scale characteristic means that the scale applies to all terms within a grade (i.e., fall, winter, spring, and summer) and across all grades K–12. Scores in the same subject area from two or more time points are comparable because they are on the same scale. For example, grade 3 Math scores in the fall may be compared with grade 4 Math scores in the spring. The equal-interval characteristic of the RIT scale allows the measurement of growth over two or more time points (i.e., learning) by calculating the difference in scores from the same vertical scale. Thus, the vertical scale and its equal-interval units are central to measuring growth using the change in scores across multiple time points.

MAP Growth efficiently and precisely measures student achievement in Math, Reading, Language Usage, and Science. Spanish-language tests are also available in Math and Reading. The adaptive nature of the assessment produces better measurement precision throughout the range of the scale than a traditional fixed-form assessment of the same length; a fixed-form test could achieve similar precision only with a longer and more time-consuming test. MAP Growth's more-efficient test length yields more time for other learning activities than a fixed-form test allows. In addition, the specific level of measurement error is reported for each student.

With MAP Growth, both in-school and remote test administration are possible. Both are handled by a trained test proctor who is available to assist students with technical challenges and other issues that may arise during testing. A proctor is in the room with students during in-school administrations, whereas remote proctoring is facilitated through a video meeting, chat, email, and 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.

Test security procedures not only address the integrity of the test itself and the test items but also the privacy of student data. Role-based access, data encryption, and system redundancy with disparate geographic locations are among the tools used for test security. During test administration, a lockdown browser may be used to inhibit students from obtaining unauthorized help in answering test items. Proctor monitoring of the test session also encourages students to put forth independent effort on the test.

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). Some of the ways MAP Growth instantiates these characteristics include item and test design, statistical analysis, bias and sensitivity review, accommodations, and eliminating threats to validity.

MAP Growth is designed to be fair for all K–12 students. The adaptive nature of MAP Growth tests meets students where they are in their achievement and selects items of suitable difficulty. Students get about half of the items correct, regardless of their ability. This feature maximizes the information the test provides about student achievement, providing insight into what a student knows and does not know. Due to its adaptive nature, everyone experiences a test of similar length and has a comparably small amount of measurement error.

A second way that MAP Growth attains fair measurement is through development and analysis procedures. Test-development procedures include reviews of item content to ensure that items reflect diversity and inclusion and do not include information that would differentially affect the performance of students because of their race, gender, or cultural background. Item development review is followed by statistical analysis techniques specifically designed to detect unfair items. Any item that shows statistical signs of unfairness is reviewed, and the item may be removed from the pool.

Universal design and test accommodations are incorporated into MAP Growth to make the test accessible for all students. For example, embedded universal design features include volume amplification, a line reader, and zoom or screen magnification. Non-embedded features, such as a Spanish dictionary, are also available. Designated features assigned to students by trained proctors include text-to-speech, color contrast, human reader, native language translation, and a separate testing setting, among many others. A variety of accommodations are available to students with an Individualized Education Plan (IEP) or 504-defined accommodation need. These include assistive technology, extended time, a screen reader, and refreshable braille.

Students motivated to perform on tests tend to achieve better test scores than students not engaged in the test (Wise & DeMars, 2005). Proctors facilitate test sessions in a manner intended to cultivate student engagement. In addition, MAP Growth includes functionality for encouraging positive test-taking behavior. Specifically, the system can detect a lack of test-taking engagement in real time. It will attempt to re-engage unmotivated students through messages and proctor intervention. The purpose is to help students give their best effort and obtain a score that more accurately reflects their achievement in a subject.

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 who is using the information. MAP Growth reports provide information for decision-making at multiple levels of an education agency, ranging from the individual student to the state department. Individual student scores may be aggregated to the school, district, or state level to summarize performance for larger groups of students. The level of aggregation should match the level of decision-making. Classroom decisions should use aggregates of individual student scores. School-level decisions should involve classroom summary statistics. District-level decisions should involve school-level summary statistics.

A variety of MAP Growth reports are available, with each serving a different purpose. MAP Growth reports also provide aggregate score summaries at multiple levels of decision-making. The MAP Growth reports portfolio summarizes each MAP Growth report as a guide for users. 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 scores typically happen in the classroom by teachers and students. For individual students, scores provide information about what they likely know and can do at a particular point in time and the extent of their learning since the last test administration. It is a direct way for students to receive feedback on their progress in school, reflect on learning, and set goals for the next term. Teachers may use the data in a similar way to monitor student learning and set goals for their students. At the classroom level, teachers may use MAP Growth scores to reflect on their teaching practices for the term and plan for the next term. When combined with other data sources, such as formative assessment, the information supports differentiated instruction and tailored learning activities for students.

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 of the 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 provide norm- and criterion-referenced information to give context to scores and facilitate score interpretation. MAP Growth norms provide a way to compare student performance with a nationally representative sample of test takers for the fall, winter, and spring terms (NWEA, 2025). 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 Index (CGI) and Conditional Growth Percentile (CGP) are two growth measures that describe a student's performance compared with similar students with comparable instructional time and prior test scores.

MAP Growth norms use a multilevel growth model that has individual students nested within schools. The model-based norms have different interpretations for individuals and schools. Individual norms allow for a student's achievement status and growth to be compared with similar students, whereas school norms allow for mean scores from one school to be compared with average scores for similar schools. The variability of scores at the individual student level and the variability of school means are key distinctions between student- and school-level norms. The variability of growth for a group of students (i.e., a school) is smaller than the variability in growth for individual students. As such, percentiles in these two cases will be

different because each represents a different level of the model. School percentiles should be obtained from school norms and not from individual norms.

Criterion-referenced information comes in the form of predictions for summative assessments and connections to external measures. NWEA linking studies connect MAP Growth scores to state summative assessments (Hu, 2021). Each study is detailed in a report that is available to partners. Linking study results are combined with the statistical model used in the MAP Growth norms in order to predict proficiency on a state assessment using MAP Growth scores from fall, winter, or spring. The information enables teachers to determine if a student is on track for proficiency by the end of the school year. This prediction is additional information a teacher may use to plan instruction and meet learning objectives. It is also used for the implementation of state policies such as “Read by Grade 3” legislation.

Linking studies also connect MAP Growth scores to other measures, with each study serving a different purpose and intention for the use of test scores. He and Meyer (2021) conducted a study that linked MAP Growth to a national composite of state summative test scores. Results provided evidence for using MAP Growth as a universal screener. Score reports provide information on MetaMetrics’ Lexile® and Quantile® measures, which open vast reading and mathematics resources to parents and educators. NWEA college readiness benchmark information uses MAP Growth scores to predict future performance on the ACT® and SAT® (Thum, 2017; Thum & Matta, 2015). The college readiness link drives the NWEA College Explorer tool and helps students plan for higher education.

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 provide a number of advantages for test design, administration, test scaling, psychometric analysis, and score interpretation. This section explains the theoretical distinctions between classical and IRT-based methods to highlight the advantages of a purely IRT-based approach. Technical information about IRT provided in this section lays the foundation for the fundamental claims of the MAP Growth assessment program and the empirical evidence needed to support them.

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 $i = 1, \ldots, n$ students responding to a set of $j = 1, \ldots, J$ 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) score[2]—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 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 concept 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 set of items that also represents the domain. Repeating this process an infinite number of times

[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.

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, $\sigma_E^2 = E\left[SEM_i^2\right]$, 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 completely confounded in the classical model. In the hypothetical sampling framework, each

[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.

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.

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

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 Chapter 8).

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 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

[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 test for differential item functioning.

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

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. Broadly speaking, elements of the target domain increase in difficulty and cognitive complexity as grade level increases. For example, fourth-grade work is more challenging than third-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 pre-requisite skills and move the student to on-grade 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 summative assessment uses blueprints focused on end-of-year learning. Some summative assessments also focus on grade-to-grade evolution and student growth from one spring to the next spring. A limitation of a traditional summative assessment is that it provides no information about the progression of student learning during the school year and focuses strictly on grade-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 blueprints represent the content domain defined by standards in mathematics, English language arts, and science. Blueprints are grade-banded or course-specific, meaning they are specific to the grade ranges or high-school level courses referenced in the test names. The grade bands are typically K–2, 2–5, and 6+, while the courses are commonly Algebra 1, Algebra 2, Geometry, and Life Sciences. This use of grade bands and courses in blueprints helps ensure that item content is suitable for different age groups. 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, such as the domain level (e.g., Algebra for math or Informational Text for English language arts) based on the specific content standards. Instructional areas are the same for every grade in the grade band or course, but standards are specific to each grade. Grade-specific standards are mapped to each instructional area.

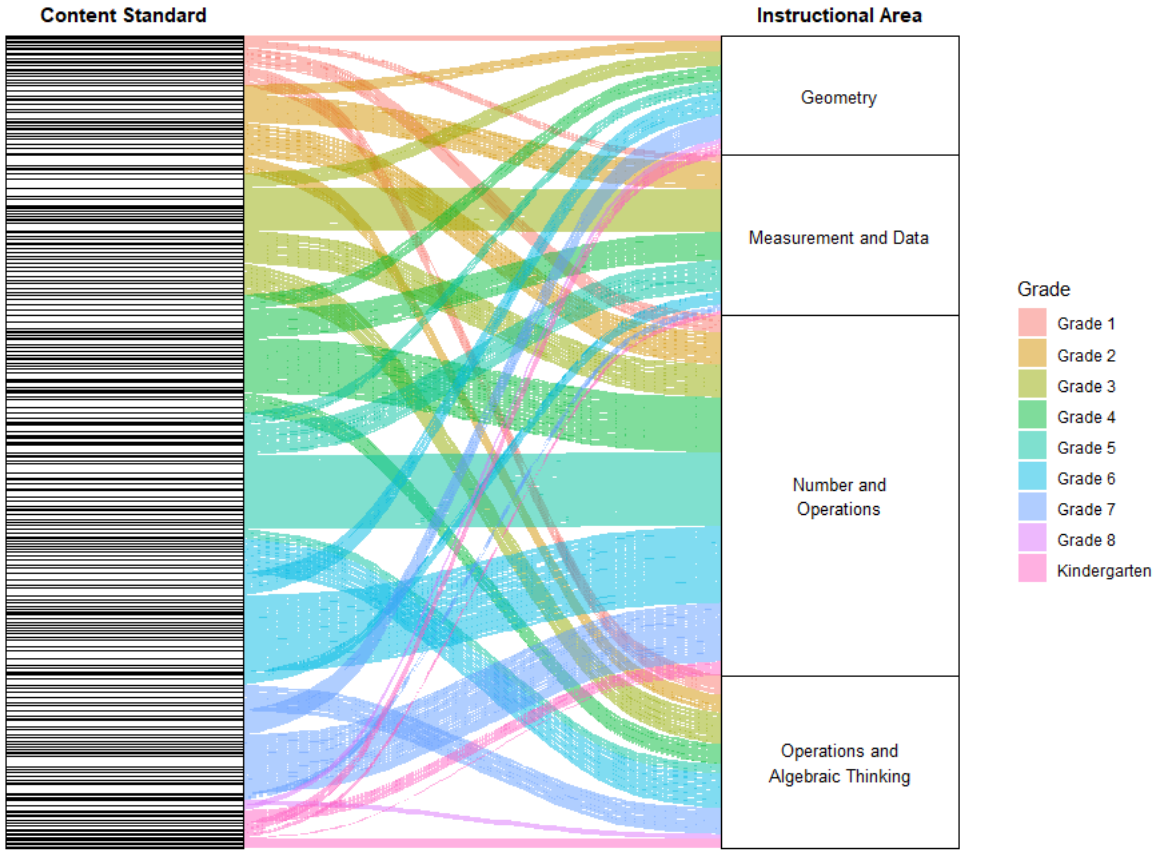
After establishing a set of blueprints, an item pool is created by choosing items from a large item bank and assigning each selected item to a content standard and instructional area. The grade level of an item is assigned according to the grade level of the standard. An item is assigned to a range of grades in cases where the standard applies to multiple grades. For example, an item aligned to a grade 9–10 Reading standard will have an item grade of both grade 9 and grade 10. The content standards represented in the blueprint go beyond the stated grade band of the test to ensure that each pool contains items for students who may be performing well-below or far-above grade level. 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 crafted for each grade-banded test for a particular subject. Items are selected from an existing bank and assigned to a content standard to create the item pool for each test.

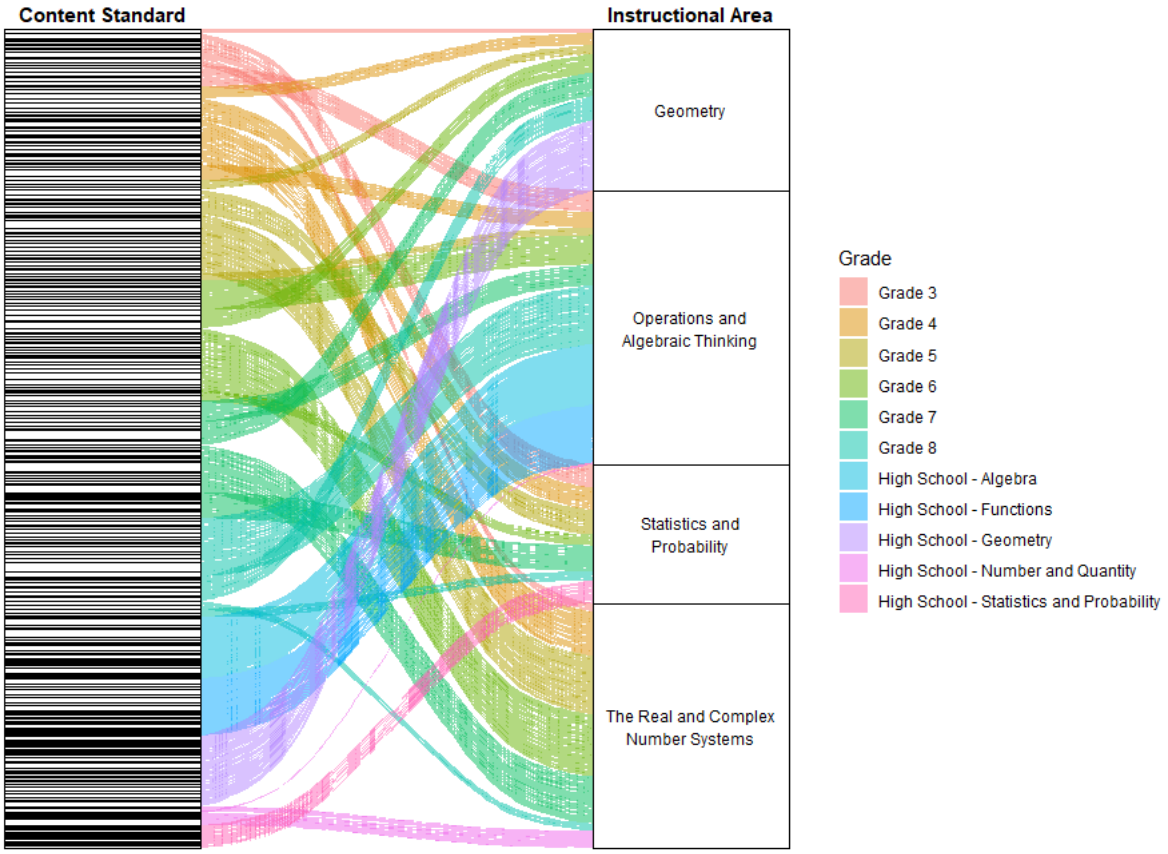
Figure 3.1 through Figure 3.5 illustrate the mapping of a state's content standards to instructional areas in proportion to the number of items for each standard and instructional area. Similar figures could be created for every MAP Growth test to show the way content standards are mapped to cross-grade instructional areas. 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. The thickness of each alluvium also represents the number of items.

Figure 3.1. Alluvial Diagram of an Item Pool for Math 2–5



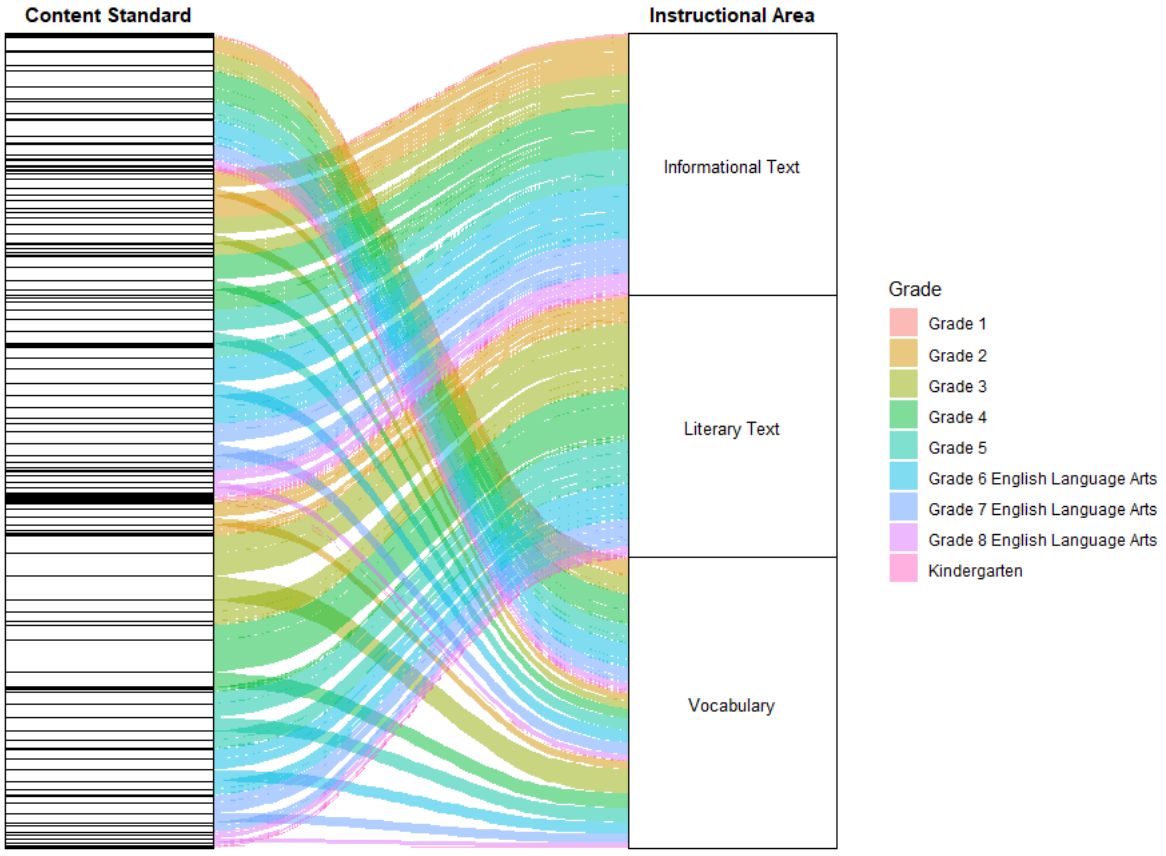
As shown in Figure 3.1, Number and Operations has the greatest number of items in the Math 2–5 item pool, while Geometry has the fewest. This design choice is based on the greater prevalence or greater emphasis of those elements in the content standards. The figure also illustrates that every grade level is represented in each instructional area. The representation is mostly uniform. However, Number and Operations has more items from grades 5 and 6 than from other grades; Operations and Algebraic Thinking also contains more items from grade 6 than from other grades.

Figure 3.2. Alluvial Diagram of an Item Pool for Math 6+



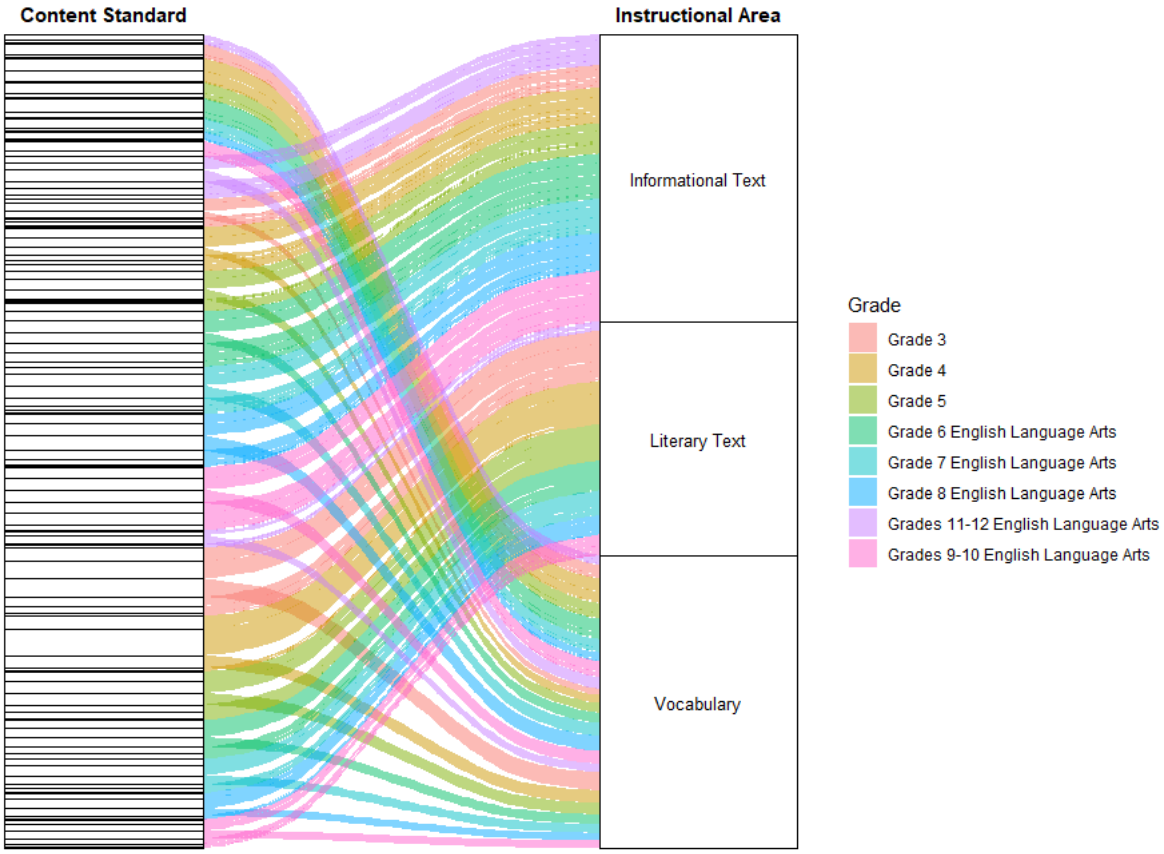
The pool for Math 6+ shown in Figure 3.2 has different instructional areas than the Math 2–5 item pool (Figure 3.1). Every grade level is represented in each instructional area. The Operations and Algebraic Thinking instructional area contains the greatest number of items, as it represents a larger number of standards. Most of those items come from grade 8 and high school. Geometry is more prevalent in the Math 6+ test than in the Math 2–5 test.

Figure 3.3. Alluvial Diagram of an Item Pool for Reading 2–5



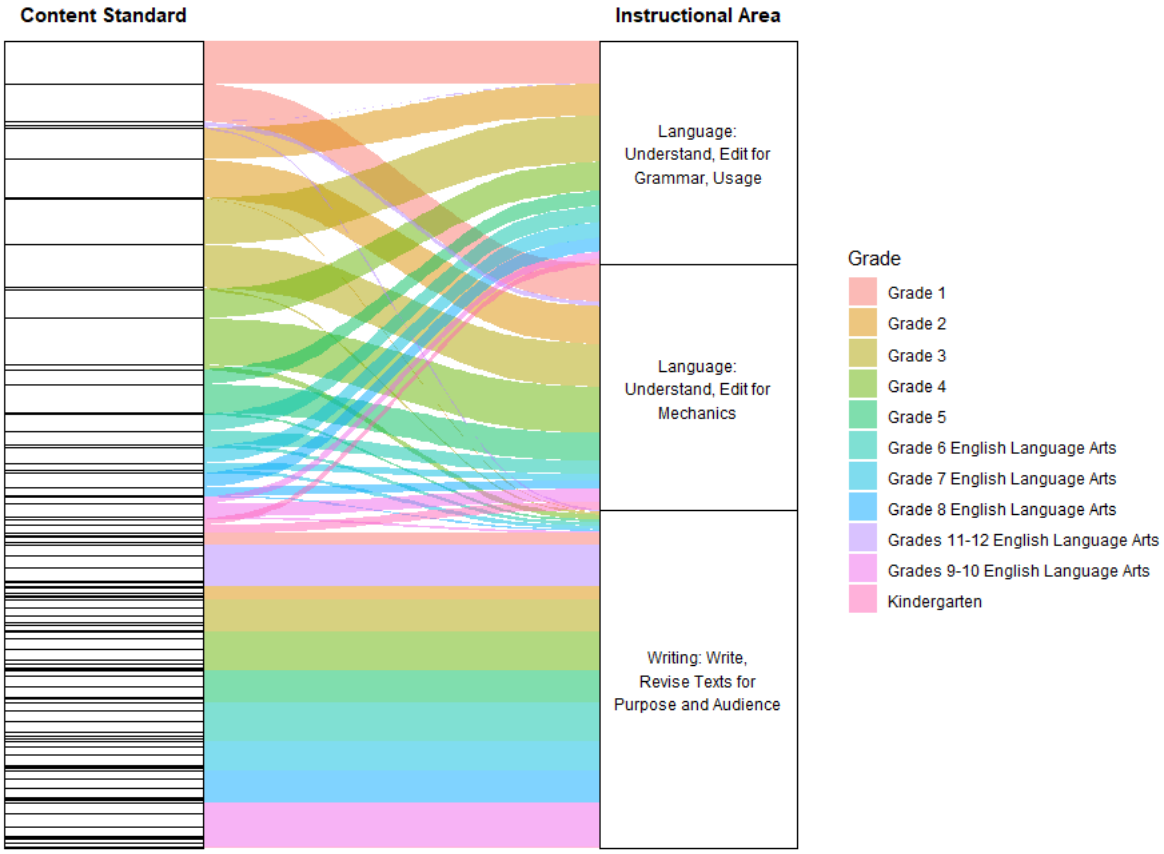
As shown in Figure 3.3, Reading 2–5 has fewer content standards and more items per standard than Math, which is evident by the larger white boxes for the content standards on the left side of the diagram. Instructional areas are represented in equal proportions, and each area includes items and standards for grades K–8. However, Vocabulary appears to be more fine-grained than the other instructional areas and thus has fewer items for each Vocabulary standard.

Figure 3.4. Alluvial Diagram of an Item Pool for Reading 6+



The same pattern of content standard to instructional area mapping is seen for the Reading 6+ test (Figure 3.4) as it was for the Reading 2–5 test (Figure 3.3).

Figure 3.5. Alluvial Diagram of an Item Pool for Language Usage 2+



Language Usage has the fewest number of standards and more items per standard, hence Figure 3.5 showing larger white boxes representing standards than were seen in either the Math or Reading test figures. The Writing instructional area has more items than the other two areas, but not by much. All grades are represented in each instructional area.

The preceding figures pertain to the tests developed for a single state. They are examples of the way state content standards are mapped to instructional areas to develop an item pool for that state. Such diagrams could be generated for every state that uses MAP Growth, though additional diagrams are not included here to avoid an excessively long technical report.

3.2. The Item Pool Is Ordered by Grade Level

Using item pools aligned to the Common Core State Standards (CCSS) for Math, Reading, and Language Usage as an example, the progression of item difficulty over terms and grades is empirical evidence that supports the ordered nature of the target domain and curriculum. Item difficulty not only spans a range of values within a grade, but the average item difficulty increases with grade level. Statistics for the 2024–2025 pool are shown in Table 3.1. Similar results are evident for other item pools used by MAP Growth because they include many of the same items, albeit with different alignments to standards. Tables for other item pools are omitted from this report for concision.

Table 3.1. Item Difficulty Summary Statistics for All Items in the CCSS Item Pool

Subject [a]	Item Grade	N	Mean	SD [b]
Math	K	841	150.07	15.24
Math	1	843	168.74	14.03
Math	2	1,052	182.63	17.94
Math	3	1,819	200.91	18.48
Math	4	1,675	214.09	18.38
Math	5	1,795	225.54	18.34
Math	6	2,132	227.67	21.32
Math	7	1,894	238.18	20.56
Math	8	961	243.95	18.06
Math	HS	2,053	259.88	19.12
Reading	K	1,072	143.84	12.04
Reading	1	1,098	159.46	12.89
Reading	2	962	176.07	12.18
Reading	3	1,768	197.19	15.55
Reading	4	1,505	206.85	12.96
Reading	5	1,246	213.58	12.91
Reading	6	1,386	216.51	11.17
Reading	7	1,051	222.72	12.46
Reading	8	781	225.08	11.38
Reading	HS	1,074	229.22	11.09
Language	K	78	170.95	12.40
Language	1	454	176.87	11.72
Language	2	415	183.61	12.53
Language	3	605	193.26	12.25
Language	4	580	200.30	12.27
Language	5	384	203.45	11.67
Language	6	348	213.13	11.63
Language	7	272	214.83	11.68
Language	8	272	220.03	11.21
Language	HS	557	223.58	9.58

Note. Some items cover a range of grades. This summary is for the lower end of the grade range.
[a] Language = Language Usage
[b] SD = standard deviation

3.3. Test Blueprints Represent the Content Standards

Blueprints guide the way an individual test event (i.e., the test experienced by an examinee) is constructed by the adaptive test engine. The blueprint examples shown in Table 3.2 list the instructional areas and other content features included in a test event. In Math, additional content features include Aspects of Rigor (AOR), which is a framework for cognitive complexity developed by Achieve (2019). In Reading, additional content features include item sets that contain a reading passage and 3–5 questions about the passage. Instructional areas and other content features are not mutually exclusive. An item may be categorized by each feature. For example, a Geometry item may also be listed as AOR: Application and AOR: Procedural. Subscores are reported for instructional areas but not for other content features. The number of

items for each content feature listed in the table is the target number of items for a test event. All test events have a maximum of 43 items, including field test items. Math, Science, and Language Usage tests have up to 3 field test items, while Reading and Spanish Reading have up to four field test items, which may include 1 field test item set. Field test item counts are not included in the numbers shown in the example blueprints in Table 3.2.

Table 3.2. Example Test Blueprints

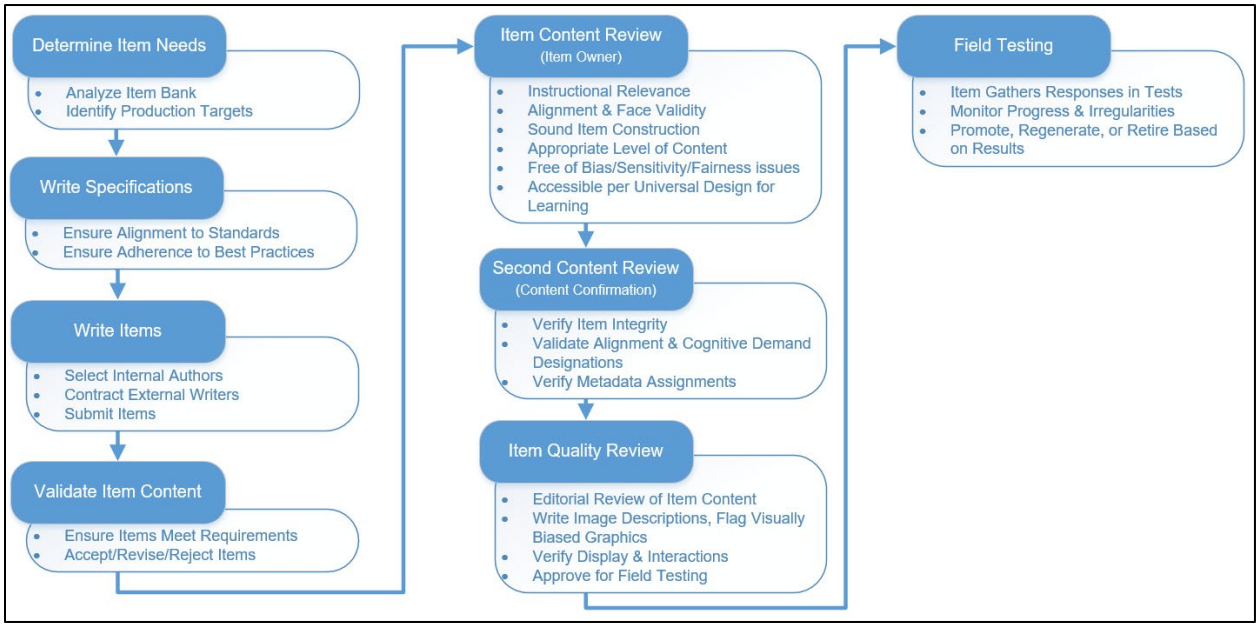
Test Name	Content Feature	Instructional Area Score	Number of Items
Math 2–5	Geometry	Yes	6
	Measurement and Data	Yes	11
	Number and Operations	Yes	13
	Operations and Algebraic Thinking	Yes	10
	AOR: Procedural	No	At least 13
	AOR: Conceptual	No	At least 16
	AOR: Application	No	At least 6
Reading 2–5	Informational Text	Yes	13
	Literary Text	Yes	13
	Vocabulary	Yes	13
	Item Set with Reading Passage	No	Up to 2
Language Usage 2+	Language: Understand, Edit for Grammar	Yes	13
	Language: Understand, Edit for Mechanics	Yes	13
	Writing: Write, Revise Texts for Purpose and Audience	Yes	13

The number of items in each instructional area in Math is chosen in proportion to their emphasis in the standards or content framework to which the test is aligned. The Science, Reading, Spanish Reading, and Language Usage tests have balanced instructional areas, with each area having the same number of items. This design choice reflects the comparable emphasis of each area in a standard set.

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 provides 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

The metadata inform whether each item is included in an item pool. For example, the "scale" field ensures that systems select only Reading items for Reading tests. For items on the Math and Science tests, metadata fields for allowable tools (e.g., ruler, protractor) and calculator (e.g., basic, scientific) determine which item tools are available during testing. Other metadata (such as grade, DOK, and item type) are used to inform item alignment to standards, development needs, and other types of internal 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. For text assets or passages, the element metadata includes readability, word count, author, and genre. Additional element data is added by permissions, 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

Growth items are aligned to multiple state content standards and frameworks using a consistent alignment philosophy that requires a direct and clearly evident connection between the skill or content targeted by the item and the skill or content described by the standard, either wholly or in a distinct and independent element of the standard. To align, items must be appropriate to the development level, difficulty level, and reading level of the standard grade and the cognitive complexity called for by the standard. Items are almost always aligned to the most granular level in a standard set's hierarchy rather than to general, over-arching standards, and are never 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 undertakes both internal and third-party alignment studies following a variety of methodologies that have provided strong validation for MAP Growth alignments across subjects and standard sets, demonstrating alignment at the item-standard level as well as via test structures. All findings from these alignment studies are reviewed by content specialists and used to inform alignment practices.

MAP Growth item-standard alignments are stored in a stable database that connects to the NWEA content management system and test-building applications so there is an authoritative record of each item's association with each standard set for a given subject. Alignments can be updated when states release new guidance about standards interpretation or if the item content is revised in a way that impacts its connection to the standard, and periodic reviews of item alignments for accuracy and consistency are conducted by content specialists.

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, as well as all materials developed for MAP Growth Reading and Language 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 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) 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
P 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 in *item sets* with multiple questions focusing on a single passage or topic or 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 (Math)

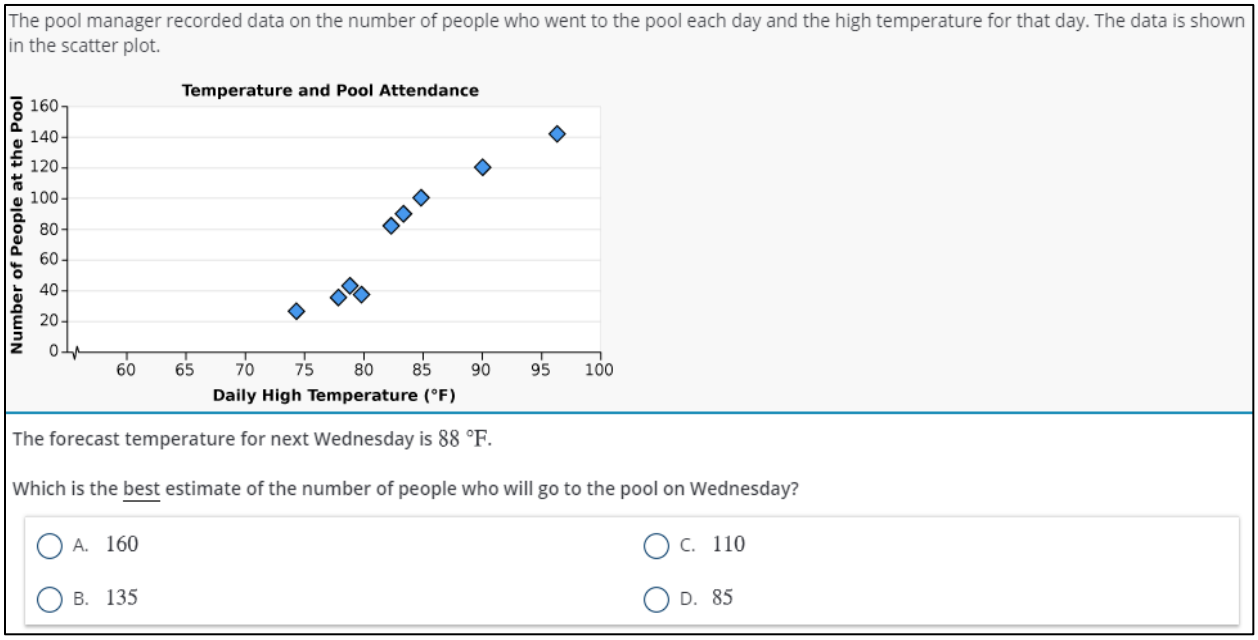


Figure 4.3. Multiselect/Multiple Select (Reading)

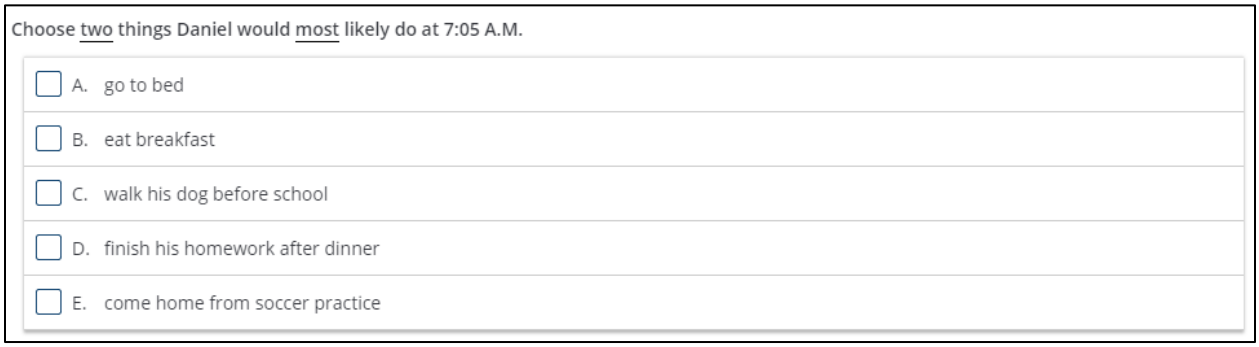


Figure 4.4. Hot Text/Selectable Text (Language Usage)

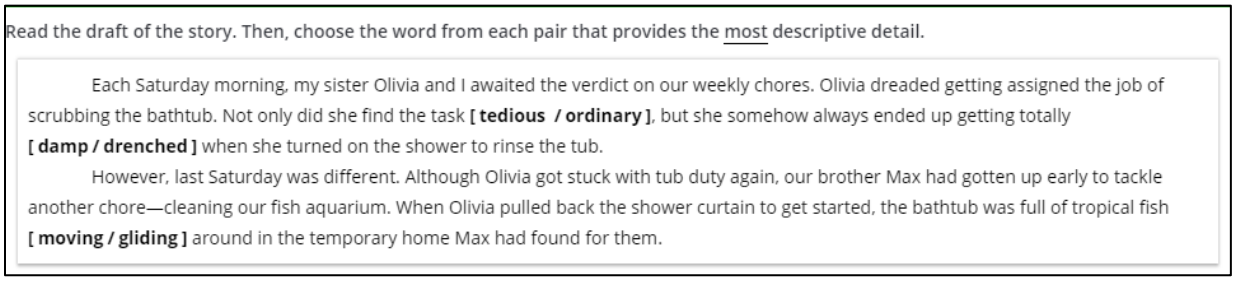
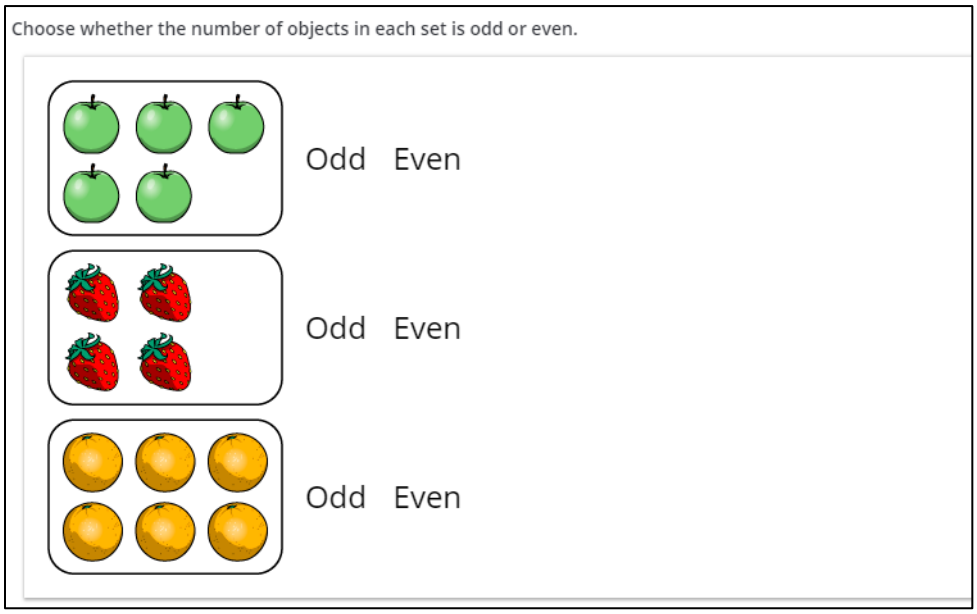


Figure 4.5. Hot Text/Selectable Text (Math)



Construction items are defined by drag-and-drop 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 (Language Usage)

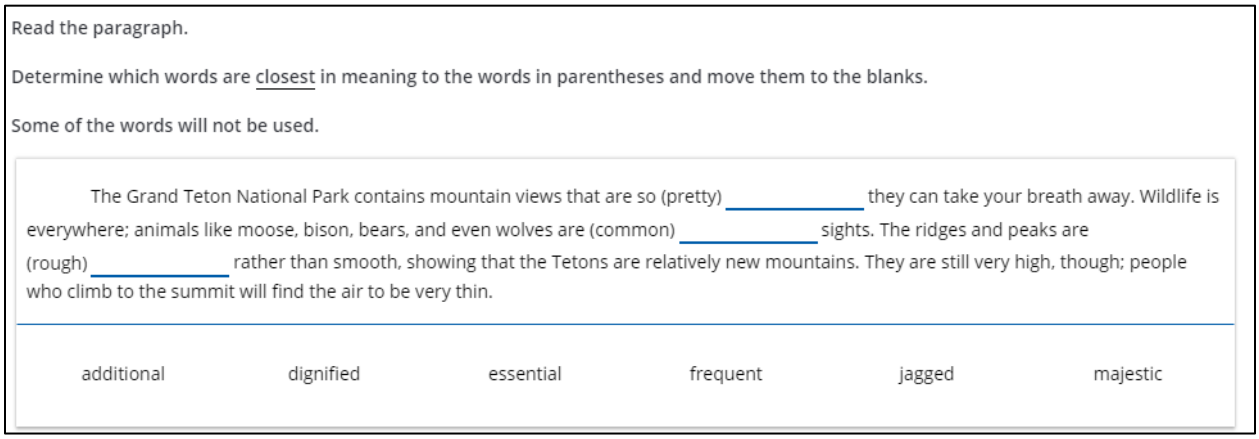
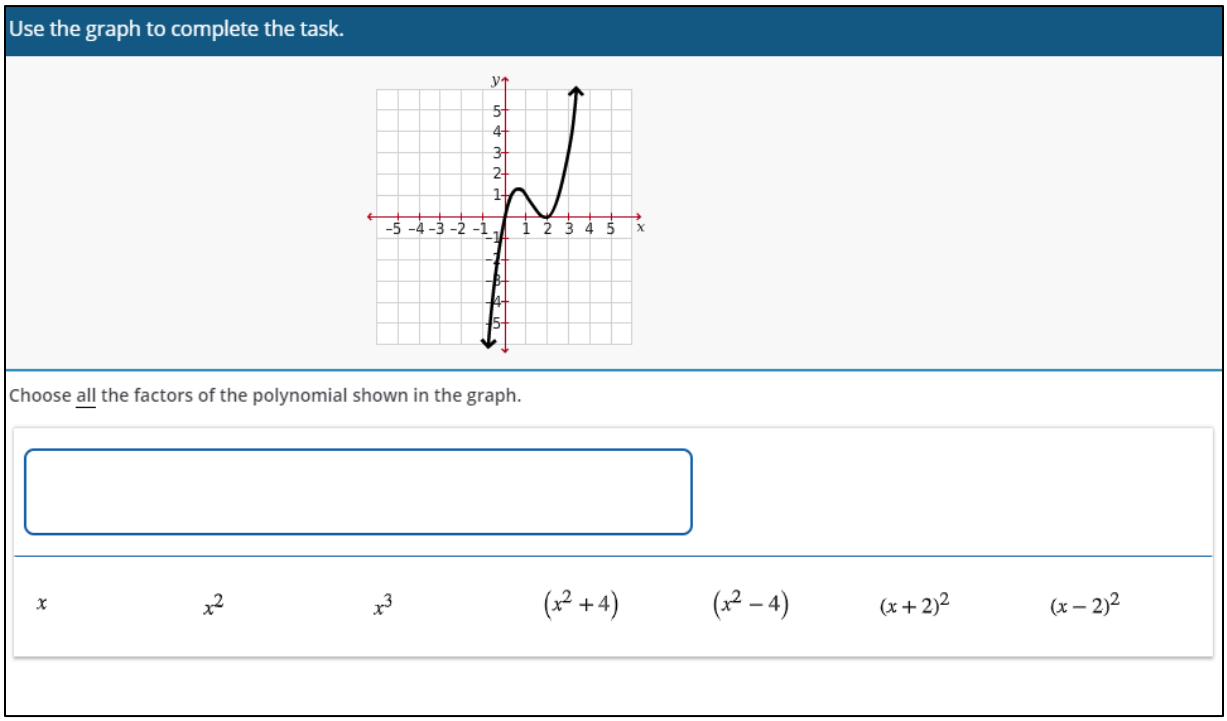
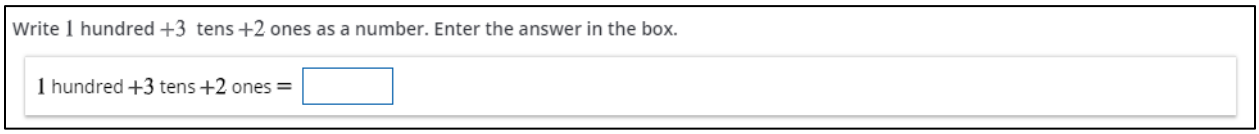


Figure 4.7. Click-and-Pop (Math)



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)



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). Typically, a reading item set has four items per passage, with the possibility of an additional field-test item.

Figure 4.9 and Figure 4.10 show examples of item sets and how a passage is combined with different selection items. The first example is a reading passage with a multiple-choice item. The next example shows the same passage and a multiselect item about it.

Figure 4.9. Item Set, Multiple-Choice (Reading)

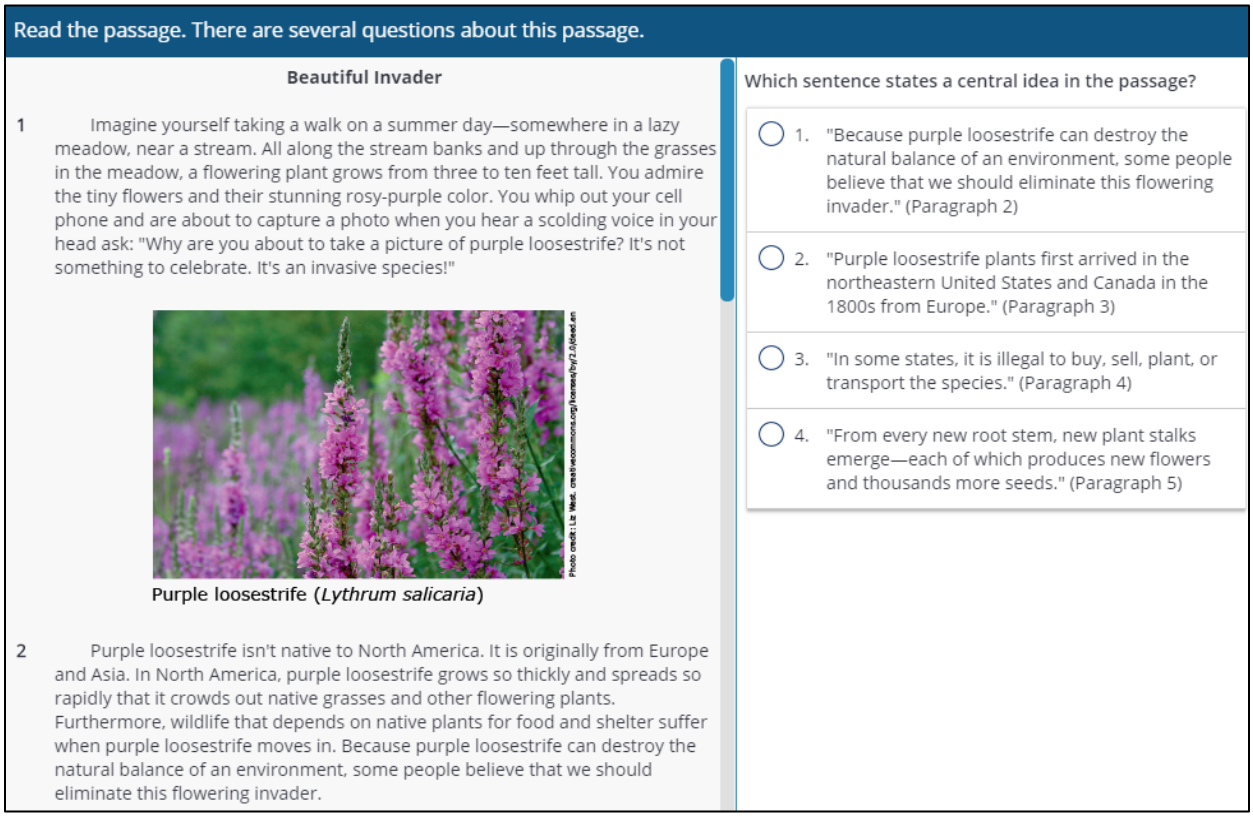
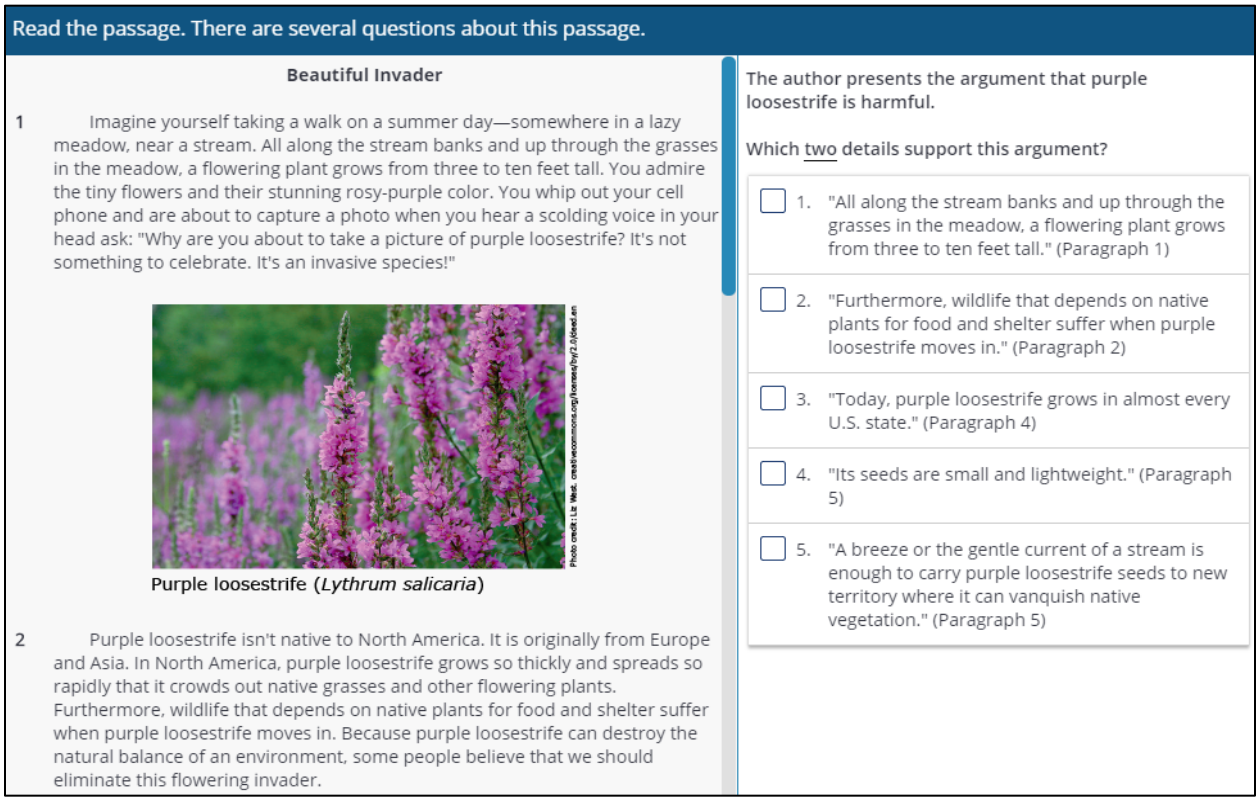


Figure 4.10. Item Set, Multiselect/Multiple Select (Reading)



An item set may also use a composite item where a stimulus is presented to students along with multiple questions. Students must answer each part correctly to get credit for the item. Figure 4.11 and Figure 4.12 shows examples of composite items and their stimuli.

Figure 4.11. Composite Item (Reading)

Read the passage and answer both questions.

1 When Marco entered the room, he thought everyone would be looking at him. After all, he was the new kid at school. His name was even written on the board at the front of the room: "Welcome, Marco!"

2 He looked around quickly, hoping to spot a friendly face. Instead, no one was looking at all. The other students were busy doing their classwork, and nobody noticed him standing there. The teacher must have stepped out of the room for a minute. Marco hesitated, then sat down at an empty desk next to a boy wearing a blue shirt.

3 The boy stopped writing and looked up at Marco. He smiled. "Hi," the boy said, "I'm Sam."

4 Marco felt relieved. "Hi," Marco answered.

5 Marco's new teacher returned and told him she would get his books for him after lunch. She seemed unsure of what to do with him in the meantime. Sam glanced around the room at the other students. Then Sam grinned and said to Marco, "I can share my book with you for now, if you want."

6 "Great," the teacher said.

7 Marco looked at his teacher and Sam and realized he had found friendly faces after all.

Which word best describes the way Sam, the boy in the blue shirt, acts?

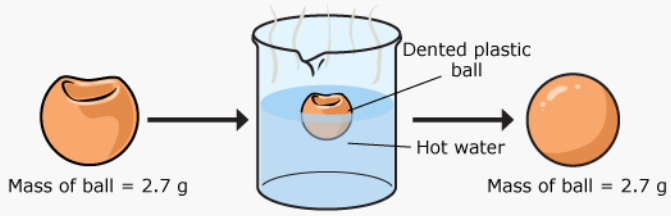
- ◯ 1. busy
- ◯ 2. careful
- ◯ 3. quiet
- ◯ 4. thoughtful

Which detail from the passage best supports your answer?

- ◯ 1. "Marco hesitated, then sat down at an empty desk next to a boy wearing a blue shirt." (Paragraph 2)
- ◯ 2. "The boy stopped writing and looked up at Marco." (Paragraph 3)
- ◯ 3. "Sam glanced around the room at the other students." (Paragraph 5)
- ◯ 4. "Then Sam grinned and said to Marco, 'I can share my book with you for now, if you want.'" (Paragraph 5)

Figure 4.12. Composite Item (Science)

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.



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.

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 2–12 assessments are untimed and take approximately 45–60 minutes per content area. MAP Growth K–2 assessments are also untimed, and students typically take 25–40 minutes per content area. MAP Growth can be administered up to four times a year (fall, winter, and spring, with 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 so they can join the testing session the proctor started.
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.

Either in-school and remote test administration are possible. Both are handled by a trained test proctor who is available to assist students with technical challenges and other issues that may arise during testing. A proctor is in the room with students during in-school administrations, whereas remote proctoring is facilitated with 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 to perform on tests, they tend to do better and the results are more likely to accurately reflect what they know and can do. In 2017, NWEA introduced the test engagement capability that detects in real-time when a student is “rapid-guessing” on items and notifies proctors so they can re-engage the student with the test. A summary of the test engagement functionality is as follows:

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

MAP Growth employs a sophisticated method for stabilizing testing accuracy when a student disengages. The average amount of time that students take to answer each unique test item is used to determine if a student has rapid-guessed when answering an item. After a student rapid-guesses one item, the difficulty of the next item locks to the same level to prevent this downward drift. After the student has rapid-guessed three items in a row, the proctor is notified and the student’s device pauses so that they can intervene and re-engage the student. The data from this test event then shows in reporting the percentage of the assessment that the student rapid-guessed and the estimated impact the disengagement could 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 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

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 information. A common criticism of test security relative to adaptive tests is that some tests do not use sufficiently large item pools to ensure that content on the test cannot be “poached” by groups of students or educators who memorize, compile, and share large numbers of items. However, well-designed, adaptive tests such as MAP Growth that draw from large item pools offer several advantages for ensuring test and item security.

The 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 for someone to anticipate and practice for the specific questions that will be encountered during a test event. Finally, students can easily be retested using a new set of items if there are questions about the integrity of their scores.

Other test security features are inherent to the administration itself. When a student logs into a test session, the test is not started and no test items are made visible to the student until the proctor has confirmed the student and activated the test session by using the proctor dashboard. In addition, item responses are not stored/cached locally. Responses are captured in real-time and stored in secure servers before presenting the next item to the student. Finally, a lockdown browser prevents students from initiating other browser sessions and having access to other content on the testing device unless they exit the test.

Furthermore, the processes and tools provided in Table 5.1 are used to ensure that the integrity of the tests are not jeopardized, thereby providing educators and students with a positive and 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 function for maintaining test security. MAP Growth uses role-based access security controls that allow partners to segregate duties in their MAP Growth accounts and grant only the amount of access needed for users to perform their jobs, as outlined in Table 5.2. This allows partners to control what actions and data individuals have access to. When planning partners' access-control strategy, MAP Growth supports granting users the least privilege to perform their work. Each role in MAP Growth has specific permissions that control levels of access to implementation, configuration, data management, testing, and reporting tasks. Each user has a unique username to which one or multiple roles can be assigned. Only certain roles can create or modify student profiles, which limits the ability to change student information. More information on NWEA 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 Off-Grade in a Principled Way

All test events aim to satisfy statistical and content requirements. In a linear test, balancing these two aspects is straightforward in that a human designer simply chooses test items and the test form is the same for everyone; the test does not change or adapt for each student. For a computer adaptive test (CAT), however, 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. The algorithm must decide on the fly how to select an item for a test to satisfy the intended statistical and content requirements. Psychometricians have developed a variety of item-selection algorithms for balancing statistical and content requirements, and each has its own strengths and weaknesses, though all aim to achieve the statistical and content requirements established in the test's design.

This chapter provides an overview of CAT algorithms in general followed by details about the MAP Growth CAT engine. Evidence is then presented to show that the MAP Growth engine is functioning as designed and delivering 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 aspects of a CAT. Maximum information item selection is a procedure where the most informative item is selected at each iteration. 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. An algorithm may achieve this goal 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. The most optimal test possible may not be achieved for everyone, but a key benefit of an adaptive test compared with a fixed-form test is that everyone will have an SEM that is no larger than the one specified in the termination rule.

The maximum information approach to item selection only accounts for a test's measurement precision. It is simply a function of person ability and item difficulty. It does not incorporate any aspect of test content outside of its relationship to item difficulty. However, to meet the test specification requirements, the item selection algorithm must be redefined to explicitly account for test content requirements.

6.1.2. Content Requirements

Content requirements are defined by test blueprints. They are not part of an item-response model and cannot be controlled through an algorithm that solely relies on maximum information item selection. Maximum information item selection must be augmented in some way, or the

optimization problem must be redefined altogether, to produce a test event that satisfies content requirements.

Heuristic methods for item selection use either a partitioning of the item pool or a mathematical function to select items and meet statistical and content requirements. Still, a limitation of all 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 it 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 used with only a few mutually exclusive content categories; however, when items have multiple and overlapping content assignments, partitioning the item pool into groups may result in empty or very sparse partitions, resulting in over exposure or item starvation (i.e., no available items).

Heuristic methods based on a mathematical function of statistical and content requirements can more easily handle items with numerous content categories. Three examples of this approach are the weighted deviations model (WDM), the weighted penalty model (WPM), and the maximum priority index (MPI). In Stocking and Swanson's (1993) weighted deviations model, the difference between the observed 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 a weighted sum, with weights indicating the importance of each content category. The item with the lowest weighted deviation is then selected. An advantage of the WDM is that it allows for both overlapping requirements and exclusion requirements—the content requirements need not 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 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 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 item's grade. Taken together, the adaptive engine uses an examinee's grade level and ability estimate to select items according to difficulty, test content, and grade level.

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 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. The weight decreases as the distance between the item grade and the student grade increases. For example, items aligned to a third-grade standard are more likely to be shown to a student in third grade than items aligned to a fourth-grade standard. Items aligned to a second-grade standard are less likely to be shown to a third grader than third-grade items but are more likely than first-grade items, and so on. The priority of selecting an item is highest when the item grade matches the student's grade; the priority decreases as the item grade differs from the student's grade. The EISA algorithm continues to include item exposure control and longitudinal constraints.

The EISA algorithm 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. The algorithm adapts off grade depending on the student's ability level. Students will see below- or above-grade items when those items are more suitable than on-grade items. A low-performing student is likely to see more below-grade items than a typical student, and a high-performing student is likely to see more above-grade 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 below-grade item will provide more information about that student's ability level. The same reasoning applies to the other end of the continuum: Above-grade items provide more information about a high-ability student than do most of the on-grade items.

The EISA algorithm 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 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 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. The curriculum experienced by these students is a mix of below-grade, on-grade, and above-grade content. Therefore, the adaptations of the EISA algorithm during the test event reflect the curriculum and instruction likely experienced by the student.

The selection of off-grade items during a test event should follow a specific pattern. Low-performing students should see more below-grade items than other students; high-performing students should see more above-grade items than other students. Items should be selected from adjacent grades more often than from more distal grades. The more extreme the level of student achievement, the more extreme the grade-level of the item. For example, an extremely low-performing student could see items from two or three grades below their enrolled grade. At the other end of the continuum, an extremely high-performing student could see items two or three grades above their enrolled grade.

6.2. Simulation Procedures and Test Publishing

With a fixed form test, every test form is designed to meet test blueprints. It is known in advance of testing if the forms adequately align with standards and meet other criteria specified in test blueprints. A CAT is different because each student takes 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 blueprints. 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.

During a simulation study, “simulees” are created from a normal distribution by generating test scores that have the same mean and standard deviation as the target student population. The generated scores are true ability scores or known examinee ability parameters. The simulation is conducted by running a CAT for these simulees. 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 examinee’s momentary ability level (momentary score) is calculated, and the momentary score is 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 the accuracy of the calculation. Other statistics are calculated to evaluate other characteristics of a CAT.

Simulation studies are conducted prior to test administration to determine if the item pool is sufficient for testing (e.g., has breadth and depth) and to evaluate that all test events meet blueprint requirements to a satisfactory degree. Conducting a simulation study prior to the operational CAT serves two main purposes:

1. To determine if the item pool is sufficient to find a feasible set of items for students at different ability levels
2. To evaluate the functioning of the engine’s item selection algorithm to ensure that the test events satisfy blueprint requirements

Simulation results are intended to demonstrate that students receive comparable representations of content with sufficient technical adequacy and infer that test scores have the same meaning across students’ test content. Table 6.1 lists the aspects of the simulation results that are evaluated and summarized in a report. The report is thoroughly reviewed against 14 or more predefined criteria and a decision—Pass, Qualified Pass, or Fail— is made about the simulation. A test is published if it results in a Pass or Qualified Pass decision. For a Qualified Pass result, a plan is developed to enhance the item pool and address weaknesses identified in the simulation. A test that fails a simulation is modified by adding items to the item pool or revising the test model. It is then submitted for another simulation.

The analysis criteria serve several purposes, each providing specific insights into the test event. Content specifications ensure that the test contains the correct number and type of items, verifying operational and field test item counts and confirming that all students receive adequate coverage across item types. Score estimation criteria assess the accuracy of score recovery,

using metrics such as root mean squared error (RMSE) and the correlations 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 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 itself marked as Pass, Flag, or Not Applicable, and the overall validation status is determined based on these results, providing a comprehensive assessment of the test's psychometric quality and operational readiness.

Table 6.1. Simulation Study Evaluation Points

Category	Evaluation Point
Content Specifications	Total number of items in the target range
	Number of operational items in the target range
	Number of field test items in the target range
	Number of items per instructional area in the target range (repeated for each instructional area)
	Percentage of on-grade items above the target minimum
	Item grade distribution aligns with student score distribution and indicates principled off-grade adaptation.
	Field test items appear in designated positions.
Score Estimation	Theta ability parameter recovered with RMSE ≤ 4
	Correlation between true and estimated scores is ≥ 0.95.
Measurement Error	Median SEM ≤ 3.5
	SEM is similar across score deciles.
	Almost all students have an SEM ≤ the maximum allowed SEM.
	SEM for an instructional area is ≤ the target SEM (repeated for each instructional area).
Adaptivity	Similarity of item difficulty and momentary ability during a test event

Simulation studies are conducted prior to publishing a test. After test administration, data from real examinees are used to evaluate the extent to which the CAT satisfied blueprints and met content and statistical requirements. The results are then presented in that 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, selection is done as preferences instead of hard constraints. The engine *prefers* to meet the targets, but it will not *require* that they be met. This avoids issues with item starvation where the item pool is depleted for some examinees when requirements are imposed and no items exist. Because of the use of preferences, test events based on blueprint targets have a range of acceptable values.

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 from one blueprint to another. A chi-square goodness-of-fit test and corresponding effect

size calculations were performed for each examinee to determine the degree to which the test event satisfied test blueprints. The statistic is shown in Equation 5.

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

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.

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

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 offered 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., Reading 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 at (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., Math 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 each examinees in each effect size bin was then tabulated. The results in Table 6.2 show that a substantial majority of test events in 2024–2025 had negligible

effect sizes in Math, Language Usage, and Science, where the percentages of negligible values ranged from 89% (Winter Language Usage) to 97% (Fall Science). The remaining percentages of test events were located in the small effect size bin.

Reading was noticeably different, where the percentage of negligible effect sizes was 51% or 52%. The next largest percentages of Reading test events (42% to 43%) were in the small effect size range. The primary reason for the larger effect sizes in Reading is the use of item sets. A set of items associated with a passage is selected collectively, with four items per set. Some items sets may balance instructional areas, whereas others may prioritize certain ones; in an extreme case, an item set would represent a single instructional area. As a result, the selection of an item set can offset the instructional area counts. Thus, observed and expected item counts per instructional area in Reading are expected to be more discrepant than in other subjects.

Across all subjects, the percentage of effect sizes in each bin is stable over the year. Math, Reading, and Language Usage are within 3 percentage points over terms. Science varies by 6 percentage points by term. Taken together, results indicate that test events meet blueprint specifications every term.

Table 6.2. Percentage of Test Events in Each Range of Effect Size by Term

Subject	Effect Size Bin	Test Events by Term (%)		
		Fall	Winter	Spring
Math	Negligible (< 0.10)	96	96	95
Math	Small (0.10–0.19)	4	4	5
Math	Medium (0.20–0.39)	0	0	0
Math	Large (0.40–0.59)	0	0	0
Math	Very large (≥ 0.60)	0	0	0
Reading	Negligible (< 0.10)	51	51	52
Reading	Small (0.10–0.19)	43	43	43
Reading	Medium (0.20–0.39)	6	6	6
Reading	Large (0.40–0.59)	0	0	0
Reading	Very large (≥ 0.60)	0	0	0
Language	Negligible (<0.10)	90	89	92
Language	Small (0.10–0.19)	10	10	8
Language	Medium (0.20–0.39)	0	0	0
Language	Large (0.40–0.59)	0	0	0
Language	Very large (≥0.60)	0	0	0
Science	Negligible (< 0.10)	97	96	91
Science	Small (0.10–0.19)	3	4	9
Science	Medium (0.20–0.39)	0	0	0
Science	Large (0.40–0.59)	0	0	0
Science	Very large (≥ 0.60)	0	0	0

Note. Language = Language Usage

6.4. Test Events Adapt Off-Grade in a Principled Way

The EISA algorithm prioritizes on-grade items and allows for the selection of off-grade items when needed to appropriately adapt item difficulty to a student's ability level. This should

produce an expected pattern of results where typical students will see a large majority of on-grade items, very low-performing students will see more below-grade items, and very high-performing students will see more above-grade items. Moreover, low-performing students will see more below-grade items than above-grade items, whereas high-performing students will see more above-grade items than below-grade items.

The justification for these expected patterns is that typical students are likely to experience mostly on-grade instruction that is derived from on-grade content standards; however, students who are very low performing are more likely to experience more below-grade instruction (e.g., intensive intervention) than typical students. Furthermore, the lower the student's achievement, the more the student is likely to experience off-grade instruction (intensive intervention) than on-grade instruction. The justification is the same at the other end of the spectrum: Very high-achieving students are likely to be placed in gifted education programs or AP courses that are above the curriculum for their enrolled grade. The MAP Growth adaptive test engine is designed to account for the grade-level content standards *and* the type of instruction or opportunity to learn a student is likely to experience in school. Most grade-specific interim assessments or high-stakes summative assessments simply focus on grade-level content standards and disregard the type of off-grade instruction some students experience, such as intensive intervention, gifted education, and AP courses. The MAP Growth engine accounts for both content standards *and* the type of instruction students are likely to experience.

Table 6.3 shows the percentage of items from the 2024–2025 administration that are below, on, and above grade by each grade and score decile. These examples of grade distribution results are taken from the Common Core State Standards–aligned MAP Growth tests for Math and Reading across grades 3–8 (complete tables for every set of standards in every subject are far too extensive to include in this report). Overall, the observed patterns are highly consistent with the intended design of the EISA algorithm, which prioritizes on-grade content while adaptively incorporating off-grade items to align with students' performance levels and their likely opportunity to learn.

6.4.1. Math

In Math, results from grades 3–8 show a clear and orderly progression across achievement deciles within each grade. Students in the middle of the achievement distribution consistently received a strong majority of on-grade items. For example, in grades 4 and 6, students in approximately the fourth through seventh deciles typically received between about 82% and 97% on-grade items. Even in grades where more off-grade targeting occurs at the extremes, middle-decile students still received predominantly on-grade content.

For students in the lowest deciles, the percentages of below-grade items increased substantially and, in many cases, exceeded the percentages of on-grade items. This pattern was especially pronounced in the upper grades. For instance, grade 7 and grade 8 students in the first decile received large majorities of below-grade items, while the proportion of above-grade items remained minimal. Importantly, across all grades, low-performing students consistently received more below-grade than above-grade items, often by a wide margin. This pattern reflects the intended downward adaptation of the engine and is consistent with the instructional reality that very low-performing students are more likely to receive intensive intervention aligned to below-grade standards.

At the upper end of the distribution, the pattern reverses. Students in the highest deciles received progressively larger proportions of above-grade items. In most grades, students in the

tenth decile received approximately 30% to over 40% above-grade items, while the proportion of below-grade items remained comparatively small. Although on-grade items generally continued to represent a majority of the assessment content, there was a clear and substantial upward shift in item grade level as achievement increased. Across all grades, high-performing students consistently received more above-grade than below-grade items, providing strong evidence that the algorithm is appropriately targeting advanced content for students likely to experience accelerated or enriched instruction.

6.4.2. Reading

The Reading results demonstrate the same fundamental pattern, though the shifts toward off-grade content at the extremes are even more pronounced than in Math. Students in the middle deciles received a large majority of on-grade items across grades 3–8. In many cases, students in the central deciles received between roughly 70% and 90% on-grade content, indicating that the assessment strongly prioritizes grade-level standards for typical readers.

Among students in the lowest deciles, the percentages of below-grade items were substantial and often dominant. In multiple grades, first-decile students received approximately 80% or more below-grade items, with very small proportions of above-grade content. Across all grades, low-performing students received considerably more below-grade than above-grade items, demonstrating clear downward adaptation. This pattern is consistent with the expectation that students with very low reading achievement are more likely to receive foundational or RTI-driven instruction that includes content below their enrolled grade.

At the highest deciles, Reading shows especially strong upward adaptation. Students in the tenth decile frequently received a majority of above-grade items, with percentages in some grades exceeding 60% or even 70%. In all cases, the proportion of above-grade items substantially exceeded the proportion of below-grade items for high-performing students. This pattern suggests that the adaptive engine is particularly responsive to high achievement in Reading, reflecting the likelihood that advanced readers are placed in gifted, accelerated, or otherwise enriched instructional settings.

Overall, the observed distributions align closely with the theoretical expectations underlying the EISA algorithm. Typical students receive predominantly on-grade content, while very low-performing students receive substantially more below-grade than above-grade items, and very high-performing students receive substantially more above-grade than below-grade items. These findings provide strong empirical support that the MAP Growth adaptive engine is functioning as intended, balancing adherence to grade-level content standards with sensitivity to students' performance levels and corresponding opportunities to learn.

Table 6.3. Percentages of Items Below, On, and Above Student Grade Level

Subject	Student Grade	Decile	N	Percentage		
				Below Grade	On Grade	Above Grade
Math	3	1	394,485	36.71	60.75	2.54
Math	3	2	394,485	6.47	91.75	1.77
Math	3	3	394,484	3.06	95.06	1.88
Math	3	4	394,484	2.06	94.24	3.70
Math	3	5	394,484	2.01	91.40	6.59
Math	3	6	394,484	2.16	86.74	11.10
Math	3	7	394,484	2.30	82.06	15.65
Math	3	8	394,484	1.69	77.45	20.87
Math	3	9	394,484	1.21	72.34	26.45
Math	3	10	394,484	1.36	56.72	41.92
Math	4	1	388,615	57.41	36.96	5.63
Math	4	2	388,615	22.74	71.97	5.29
Math	4	3	388,615	11.05	84.90	4.05
Math	4	4	388,614	5.47	90.49	4.04
Math	4	5	388,614	2.82	93.45	3.74
Math	4	6	388,614	1.13	95.71	3.16
Math	4	7	388,614	0.37	96.62	3.00
Math	4	8	388,614	0.25	94.96	4.80
Math	4	9	388,614	0.73	88.06	11.21
Math	4	10	388,614	2.48	63.14	34.39
Math	5	1	394,219	55.98	27.27	16.75
Math	5	2	394,219	35.73	47.04	17.22
Math	5	3	394,219	24.70	59.68	15.62
Math	5	4	394,218	15.99	71.29	12.72
Math	5	5	394,218	10.30	79.00	10.70
Math	5	6	394,218	6.82	83.84	9.34
Math	5	7	394,218	6.37	84.69	8.94
Math	5	8	394,218	6.54	82.85	10.61
Math	5	9	394,218	6.67	78.10	15.24
Math	5	10	394,218	4.99	63.13	31.88
Math	6	1	402,589	51.56	43.47	4.97
Math	6	2	402,589	31.69	60.35	7.95
Math	6	3	402,589	18.72	72.24	9.05
Math	6	4	402,589	11.23	81.63	7.14
Math	6	5	402,589	7.97	86.79	5.24
Math	6	6	402,588	5.57	90.04	4.39
Math	6	7	402,588	3.93	91.90	4.17
Math	6	8	402,588	3.34	91.45	5.21
Math	6	9	402,588	3.74	85.30	10.96
Math	6	10	402,588	4.59	62.80	32.62
Math	7	1	406,075	73.22	20.93	5.85

Subject	Student Grade	Decile	N	Percentage		
				Below Grade	On Grade	Above Grade
Math	7	2	406,075	51.22	39.28	9.50
Math	7	3	406,075	38.62	51.02	10.36
Math	7	4	406,075	28.81	58.79	12.40
Math	7	5	406,075	21.23	64.91	13.86
Math	7	6	406,074	16.74	70.37	12.89
Math	7	7	406,074	12.51	76.17	11.32
Math	7	8	406,074	8.71	79.78	11.52
Math	7	9	406,074	6.86	80.56	12.59
Math	7	10	406,074	6.44	63.18	30.38
Math	8	1	381,185	79.65	15.51	4.85
Math	8	2	381,185	60.52	30.92	8.56
Math	8	3	381,185	48.34	42.83	8.84
Math	8	4	381,185	40.24	53.65	6.10
Math	8	5	381,184	33.89	61.35	4.76
Math	8	6	381,184	28.17	66.70	5.13
Math	8	7	381,184	21.86	72.27	5.87
Math	8	8	381,184	15.56	76.51	7.94
Math	8	9	381,184	9.55	75.44	15.01
Math	8	10	381,184	8.06	58.65	33.29
Reading	3	1	425,019	83.15	15.85	1.00
Reading	3	2	425,019	54.34	43.43	2.23
Reading	3	3	425,019	29.75	65.55	4.70
Reading	3	4	425,019	15.13	78.48	6.39
Reading	3	5	425,018	8.21	82.34	9.46
Reading	3	6	425,018	6.65	76.05	17.30
Reading	3	7	425,018	6.82	63.03	30.15
Reading	3	8	425,018	7.81	47.43	44.77
Reading	3	9	425,018	5.30	35.21	59.49
Reading	3	10	425,018	2.72	21.36	75.92
Reading	4	1	384,965	79.88	18.28	1.84
Reading	4	2	384,965	35.14	60.66	4.21
Reading	4	3	384,965	13.84	82.47	3.68
Reading	4	4	384,965	5.48	91.07	3.45
Reading	4	5	384,965	3.65	90.10	6.24
Reading	4	6	384,965	4.62	81.78	13.60
Reading	4	7	384,965	7.45	67.22	25.33
Reading	4	8	384,964	8.78	52.36	38.86
Reading	4	9	384,964	7.56	41.67	50.77
Reading	4	10	384,964	4.40	28.76	66.84
Reading	5	1	387,790	86.32	13.23	0.45
Reading	5	2	387,790	46.71	49.22	4.07
Reading	5	3	387,790	21.95	72.56	5.48

Subject	Student Grade	Decile	N	Percentage		
				Below Grade	On Grade	Above Grade
Reading	5	4	387,790	12.29	81.41	6.30
Reading	5	5	387,790	9.98	78.37	11.66
Reading	5	6	387,790	10.76	66.18	23.06
Reading	5	7	387,790	11.40	51.84	36.76
Reading	5	8	387,790	11.46	40.46	48.08
Reading	5	9	387,789	9.50	30.27	60.23
Reading	5	10	387,789	6.73	22.35	70.92
Reading	6	1	396,895	79.81	16.58	3.61
Reading	6	2	396,895	47.76	46.54	5.70
Reading	6	3	396,895	27.78	67.52	4.70
Reading	6	4	396,895	17.50	77.27	5.23
Reading	6	5	396,895	10.41	82.48	7.11
Reading	6	6	396,895	6.41	82.97	10.63
Reading	6	7	396,895	4.25	79.35	16.40
Reading	6	8	396,895	3.67	71.32	25.01
Reading	6	9	396,895	4.53	57.47	38.00
Reading	6	10	396,895	5.02	33.91	61.07
Reading	7	1	401,123	82.22	16.63	1.15
Reading	7	2	401,123	57.39	38.19	4.42
Reading	7	3	401,123	36.95	55.96	7.09
Reading	7	4	401,123	23.98	67.43	8.60
Reading	7	5	401,123	16.51	73.39	10.10
Reading	7	6	401,123	13.27	74.34	12.39
Reading	7	7	401,123	11.81	70.40	17.79
Reading	7	8	401,123	11.80	61.86	26.34
Reading	7	9	401,123	12.04	50.93	37.03
Reading	7	10	401,122	9.07	36.28	54.65
Reading	8	1	384,813	85.91	10.79	3.30
Reading	8	2	384,813	58.91	31.27	9.82
Reading	8	3	384,813	41.74	48.55	9.72
Reading	8	4	384,813	27.71	63.32	8.97
Reading	8	5	384,813	17.58	72.89	9.52
Reading	8	6	384,813	12.86	74.42	12.72
Reading	8	7	384,813	12.41	68.80	18.79
Reading	8	8	384,813	13.50	60.00	26.49
Reading	8	9	384,812	13.95	51.38	34.67
Reading	8	10	384,812	13.63	39.30	47.06

6.5. 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 Chapter 10) 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 allows scores from one time point to be compared with scores from another time point. Vertical scaling is a requirement for measuring student growth (i.e., learning). MAP Growth Math and Reading assessments use a vertical scale that spans fall, winter, spring, and summer terms across grades K–12. Both MAP Growth Language Usage and Science tests cover grades 2–12 for the same four terms. Each subject has its own vertical scale. A vertical scale allows all items in the pool to be on the same scale regardless of term or grade. MAP Growth's vertical scale for each subject originates with the NWEA Levels Tests (Ingebo, 1997). Although test content has evolved to reflect modern content standards, the scale remains the same. New items are calibrated to the vertical scale for a subject on a continuous basis. The pool is replenished over time as existing items that drift from the scale are updated with new difficulty estimates and items that reflect obsolete content are retired.

Scaling is accomplished using the Rasch model (Rasch, 1980). MAP Growth person ability and item difficulty parameters are on the RIT scale (i.e., **R**asch un**IT** 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.

Statistics were computed for every operational item used during the 2024–2025 test administration. Across all terms, Math items demonstrate a well-balanced difficulty distribution, with mean *p* values increasing slightly from fall (0.48) to winter and spring (0.51), indicating a modest increase in the proportion of correct responses over time, as shown in Table 7.1. Most items fall in the 0.4–0.7 *p*-value range, suggesting appropriate targeting for the tested population. Correspondingly, Table 7.2 shows that Math items present strong and stable discrimination, with mean point-biserial correlations ranging from 0.32 to 0.33 and the vast majority of items clustered in the 0.2–0.4 and 0.4–0.6 ranges. The absence of negative or near-zero discrimination values indicates that Math items consistently differentiate between lower- and higher-performing students while maintaining appropriate difficulty.

Reading items show a progressive increase in average difficulty from fall (mean *p* value = 0.47) to spring (mean *p* value = 0.51), with a broad distribution that includes a small proportion of very difficult items. Over time, items increasingly concentrate in the 0.5–0.8 *p*-value range, reflecting improved alignment with student ability. Reading also exhibits consistently strong discrimination,

with mean point-biserial correlations of 0.32 across all terms and a narrowing distribution of values by spring. The lack of negatively discriminating items supports the conclusion that Reading items reliably measure the intended construct and effectively distinguish performance levels.

Language Usage assessments are characterized by relatively higher average *p* values (means of 0.51–0.53), indicating generally easier items compared with other subjects. Item difficulty distributions are tightly clustered, with most items in the 0.5–0.8 range and very few extreme values. Discrimination is consistent and acceptable, with mean point-biserial correlations of 0.31 across all terms and minimal variability. The stable combination of moderate difficulty and solid discrimination suggests a well-calibrated item pool with limited need for revision.

Science items show moderate difficulty, with mean *p* values increasing from 0.47 in fall to 0.50 in spring, indicating a modest increase in the observed proportion correct over time. Early administrations include a slightly higher proportion of items in lower *p*-value ranges, while later terms show stronger clustering in the 0.4–0.7 range. Mean point-biserial correlations increase from 0.30 to 0.31, indicating adequate and improving discrimination. Although Science includes marginally more lower-discriminating items than other subjects, the overall distribution remains strongly positive, supporting effective differentiation across the score scale.

Table 7.1. Summary of Item *P* Values

Subject [a]	Term	N	Mean	SD [b]	*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]
Math	Fall	17,693	0.48	0.10	0.03	0.33	3.13	16.33	37.83	32.43	8.82	1.03	0.07	0
Math	Winter	17,999	0.51	0.10	0.02	0.32	2.52	12.35	31.36	35.31	15.58	2.38	0.16	0.01
Math	Spring	18,546	0.51	0.10	0.03	0.27	1.97	10.67	31.56	36.62	16.04	2.60	0.24	0.01
Reading	Fall	11,070	0.47	0.10	0.21	1.23	3.84	14.40	41.43	29.84	7.51	1.17	0.35	0.03
Reading	Winter	11,057	0.50	0.11	0.21	1.11	3.11	10.92	35.56	33.02	12.29	3.12	0.60	0.06
Reading	Spring	11,091	0.51	0.12	0.21	0.98	2.80	10.05	33.41	33.60	13.69	4.44	0.75	0.06
Language	Fall	4,454	0.51	0.08	0	0.04	0.49	8.17	37.23	39.47	13.29	1.30	0	0
Language	Winter	4,402	0.53	0.09	0	0.07	0.34	6.41	31.78	42.66	16.33	2.41	0	0
Language	Spring	4,458	0.53	0.09	0	0.07	0.25	5.36	30.22	43.29	17.72	3.10	0	0
Science	Fall	5,129	0.47	0.09	0.08	0.39	3.22	16.55	40.12	32.87	6.38	0.35	0.04	0
Science	Winter	5,102	0.49	0.09	0.08	0.49	2.61	13.29	36.95	36.65	8.82	1.04	0.06	0.02
Science	Spring	5,189	0.50	0.09	0.08	0.44	2.56	10.87	36.31	38.08	10.23	1.35	0.08	0

[a] Language = Language Usage
[b] SD = standard deviation

Table 7.2. Summary of Item Point-Biserial Correlations

Subject [a]	Term	N	Mean	SD [b]	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]
Math	Fall	17,693	0.32	0.06	0	0	0	0	0.02	3.71	86.68	9.59	0	0
Math	Winter	17,999	0.33	0.06	0	0	0	0	0.01	3.23	86.19	10.57	0	0
Math	Spring	18,546	0.32	0.06	0	0	0	0	0.02	3.18	88.14	8.66	0	0
Reading	Fall	11,070	0.32	0.06	0	0	0	0	0	2.80	89.05	8.15	0	0
Reading	Winter	11,057	0.32	0.06	0	0	0	0	0	2.30	90.17	7.53	0	0
Reading	Spring	11,091	0.32	0.06	0	0	0	0	0	2.33	91.18	6.49	0	0
Language	Fall	4,454	0.31	0.06	0	0	0	0	0	3.88	92.34	3.77	0	0
Language	Winter	4,402	0.31	0.06	0	0	0	0	0.02	3.54	91.44	5	0	0
Language	Spring	4,458	0.31	0.06	0	0	0	0	0.02	3.48	92.46	4.04	0	0
Science	Fall	5,129	0.30	0.06	0	0	0	0	0.02	4.82	90.43	4.74	0	0
Science	Winter	5,102	0.31	0.06	0	0	0	0	0.02	4.70	89.69	5.59	0	0
Science	Spring	5,189	0.31	0.06	0	0	0	0	0.04	4.64	89.11	6.21	0	0

[a] Language = Language Usage
[b] SD = standard deviation

7.2. Item Fit

Item OUTFIT is a type of fit statistic calculated on items administered to students. The expected value of the OUTFIT statistic is 1. Values close to 1 are considered good-fitting items, with acceptable values ranging between 0.5 and 1.5; misfitting items have values outside this range.

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 for the 2024–2025 school year. Almost all items were within the acceptable range, though some did fall outside. Across all subjects, the boxes indicate that 50% of fit statistics fell between about 0.9 and 1.1, which is well within the acceptable range. Item fit statistics showed more variability in Math than in other subjects, as indicated by taller boxes in the top-right panel of the figure.

A limitation of the information in Figure 7.1 is that fit statistics are computed for each term separately. The items showing misfit in one term may be entirely different from items showing misfit in another term, or it may be the same items misfitting in both terms. The analysis does not show if an item may be misfitting across terms. Therefore, item fit is studied longitudinally to determine whether items remain stable and maintain fit across terms.

7.2.2. Longitudinal Item Fit

Item fit statistics are also computed longitudinally for items given in fall, winter, and spring that have 250 or more responses. Each item had to be given in all three terms and administered as an operational item (i.e., not a field test item) at the time of testing. Fit statistics should be within the acceptable range of values at every term. If a value is consistently outside the acceptable range, then an updated calibration is warranted.

Table 7.3 shows the percentage of items in each subject that showed no misfit in any term or the number of terms where an item showed misfit. In Math, 94% of the items showed no misfit in any term, whereas about 4% of the items showed misfit in only one term (e.g., fall, winter, or spring). A smaller percentage showed misfit in two terms (e.g., fall and winter, fall and spring). A very small percentage of Math items (0.38%) showed misfit in all three terms.

A similar pattern is evident for Reading, Language Usage, and Science items. The vast majority of items showed no misfit. Less than 2% of the items showed misfit in one term, and less than 1% showed misfit on two or three terms.

Figure 7.1. Item OUTFIT Statistics by Subject and Term

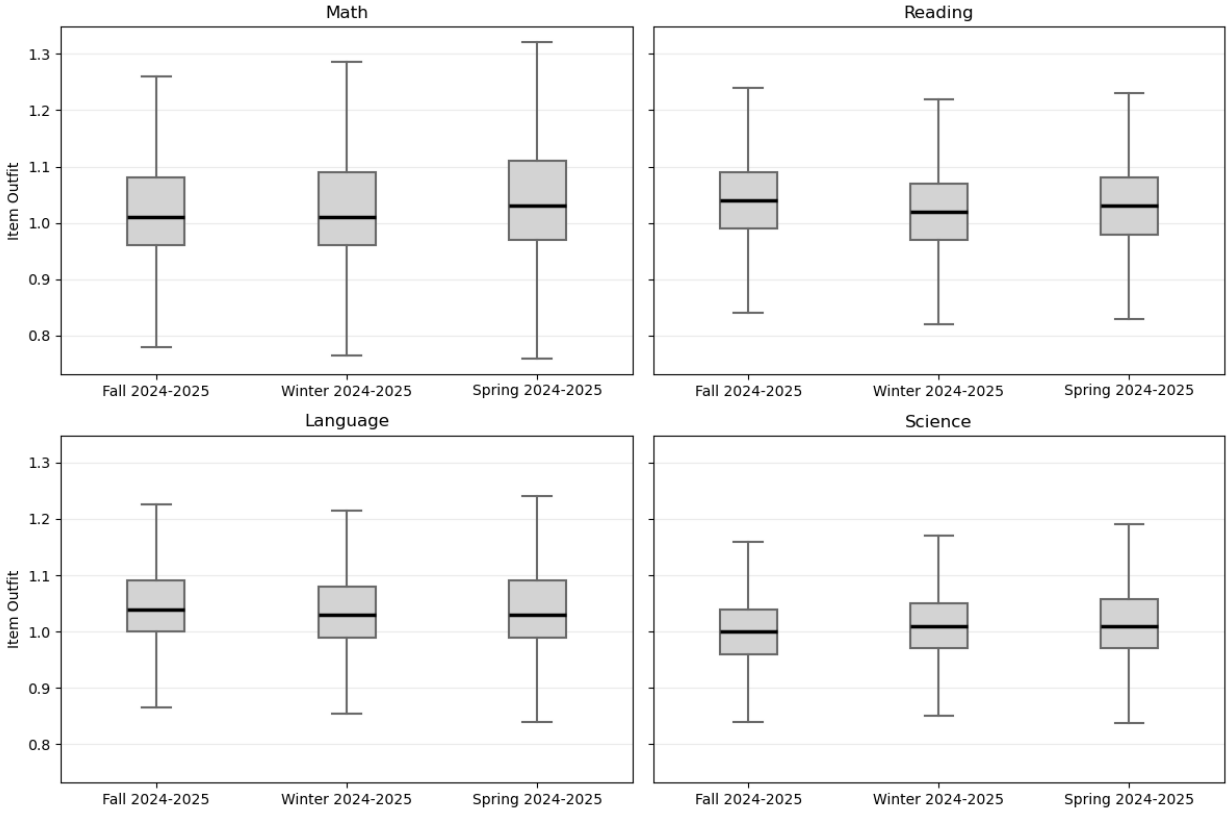


Table 7.3. Percentages of Items Misfitting Across Terms

Subject	Misfitting Terms [a]	Number of Items [b]	Percentage
Math	0	19,970	94.13
Math	1	947	4.46
Math	2	218	1.03
Math	3	80	0.38
Reading	0	11,337	98.44
Reading	1	100	0.87
Reading	2	36	0.31
Reading	3	44	0.38
Language Usage	0	4,617	98.53
Language Usage	1	65	1.39
Language Usage	2	3	0.06
Language Usage	3	1	0.02
Science	0	6,203	98.38
Science	1	80	1.27
Science	2	15	0.24
Science	3	7	0.11

[a] Number of terms where an item is misfitting.
[b] Items are counted multiple times when observed in more than one term.

These fit statistics indicate that the vast majority of items have remained stable over time. Very few showed misfit in two terms, and less than half a percent showed misfit in all three terms. The implication is that item difficulty for the vast majority of items is stable because model-data fit remains comparable over time when the item parameter estimates are held constant.

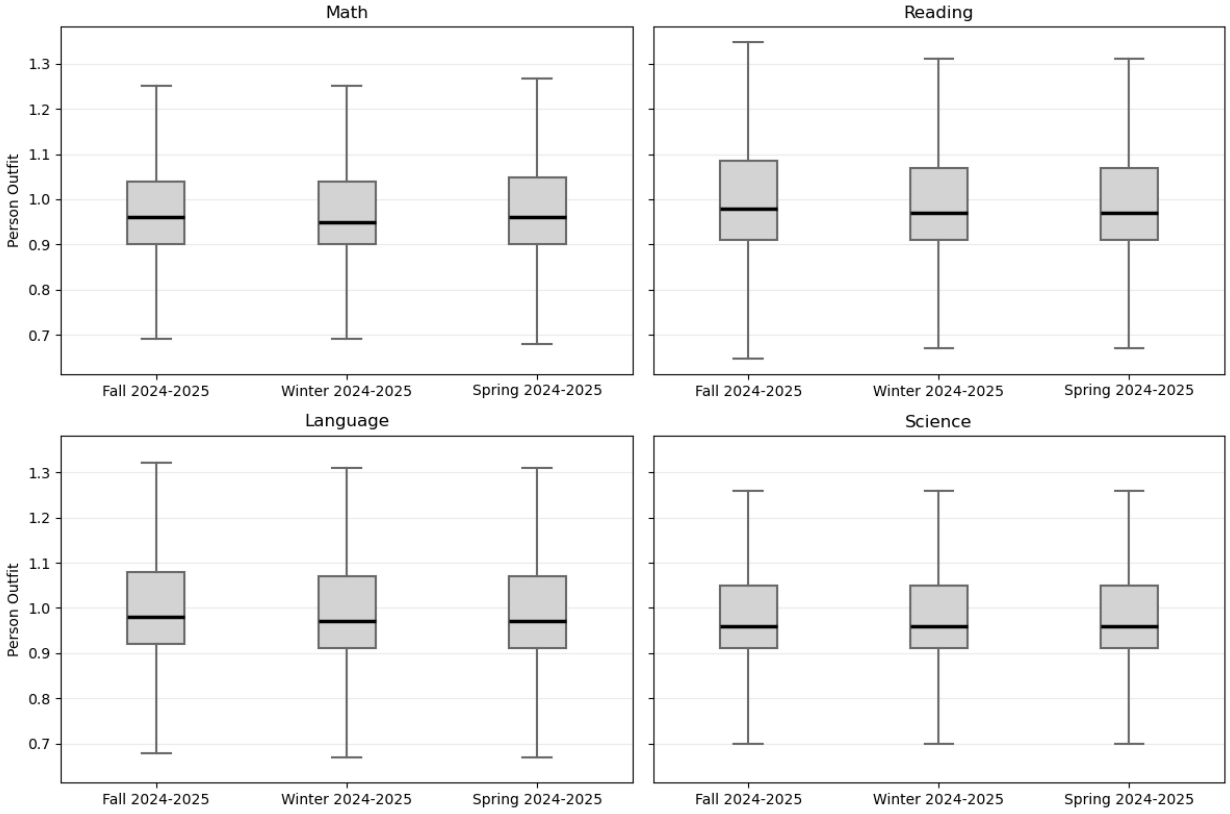
The complete distribution of OUTFIT statistics for the 2024–2025 school year is illustrated across Figure 7.1 through Figure 7.3. Horizontal lines are added to each plot at 0.5 and 1.5 to mark the range of acceptable values.

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 score fits the data, given the items completed by the examinee. It is interpreted in the same manner as item OUTFIT such that the expected value is 1 and the range of acceptable values is between 0.5 and 1.5. As shown in Figure 7.2, the median fit statistic is slightly below 1 for every subject, and 50% of the values are typically between 0.9 and 1.1 across subjects, as indicated by the box in each figure. The whiskeys in each plot extend 1.5 times the interquartile range beyond the box, and even those values are in the acceptable range of the OUTFIT statistic.

Figure 7.2. Examinee OUTFIT Statistics by Subject and Term



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

An assessment program based on grade-specific scaling cannot show the progression of achievement from one grade level to another because each grade-level test is independently scaled and disconnected from tests given in other grade levels. Inferences are limited to grade-level achievement. By contrast, an assessment program with a vertical scale can show achievement at a specific point in time and how student achievement progresses over time (i.e., growth). Examinee scores and item difficulties are on the same scale in the Rasch model. As such, the average item difficulty should also progress with grade level if the items are well-matched to the examinees.

Table 7.4 provides evidence of the vertical scale and shows that both examinee scores and item difficulty values for 2024–2025 increased on average with grade level. Student average scores are higher than average item difficulty in the lower grades. The trend reverses in higher grades. The variability of scores is similar to the variability of difficulties for each grade. The results in the table are specific to the CCSS-aligned tests and scores from students who took those tests.

Table 7.4. Comparison of Examinee RIT Scores to Item Difficulties

Subject	Grade	Students		Items	
		Mean	SD	Mean	SD
Math	K	162.97	13.26	150.07	15.24
Math	1	179.30	14.72	168.74	14.03
Math	2	191.49	16.15	182.63	17.94
Math	3	203.10	17.30	200.91	18.48

Subject	Grade	Students		Items	
		Mean	SD	Mean	SD
Math	4	213.60	18.75	214.09	18.38
Math	5	218.05	19.60	225.54	18.34
Math	6	223.03	19.65	227.67	21.32
Math	7	227.21	20.53	238.18	20.56
Math	8	232.16	21.44	243.95	18.06
Math	HS	233.16	22.95	259.88	19.12
Reading	K	156.14	13.92	143.84	12.04
Reading	1	171.69	16.04	159.46	12.89
Reading	2	184.84	17.00	176.07	12.18
Reading	3	195.19	17.88	197.19	15.55
Reading	4	203.32	17.69	206.85	12.96
Reading	5	208.91	17.28	213.58	12.91
Reading	6	213.05	16.15	216.51	11.17
Reading	7	216.72	15.94	222.72	12.46
Reading	8	220.04	15.79	225.08	11.38
Reading	HS	220.84	16.36	229.22	11.09
Language	2	187.25	16.84	183.61	12.53
Language	3	197.23	15.75	193.26	12.25
Language	4	204.95	15.53	200.30	12.27
Language	5	210.68	15.30	203.45	11.67
Language	6	213.88	15.98	213.13	11.63
Language	7	217.10	15.38	214.83	11.68
Language	8	219.69	15.87	220.03	11.21
Language	HS	217.43	18.93	223.58	9.58

Note. Language = Language Usage; 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.

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

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 test event is designed to have a target CSEM of about 3.5 RIT points. For a given test, the target CSEM is the same for all students, and the observed SEM should be similar for all students regardless of their ability. Figure 7.3 displays boxplots of CSEMs for students taking MAP Growth during the 2024–2025 school year who were grouped into deciles. The plots show that the CSEM is fairly constant across deciles, which is expected for an adaptive test. An exception is the first decile, where CSEM values are noticeably larger than for other deciles. The larger values indicate that fewer items are available for very low-performing students, resulting in less adaptivity and higher CSEM values than in other deciles.

Table 7.5 shows numerical summaries of the CSEM value plotted in Figure 7.3, but the information is limited to middle and extreme deciles (i.e., 1st, 5th, and 10th decile). The average CSEM appears to be similar from one term to the next for the same grade and subject. CSEM values are slightly larger for Reading than for other subjects, and the values increase slightly by grade for Math. The lowest values are observed for students in the middle decile, while slightly larger values are observed for students at the extreme deciles. As shown in Table 7.6, when CSEM values are analyzed by subject, demographic, term, and decile group, values are similar across all facets *except* decile group. That is, the decile group appears to affect the CSEM with slightly larger values in extreme deciles, whereas the mean CSEM is similar across all other factors (e.g., demographic and term) more than any other facet. In cases where a pattern is observed, the difference is a fraction of a RIT point.

Figure 7.3. Boxplots of CSEM by Subject and RIT Score Decile

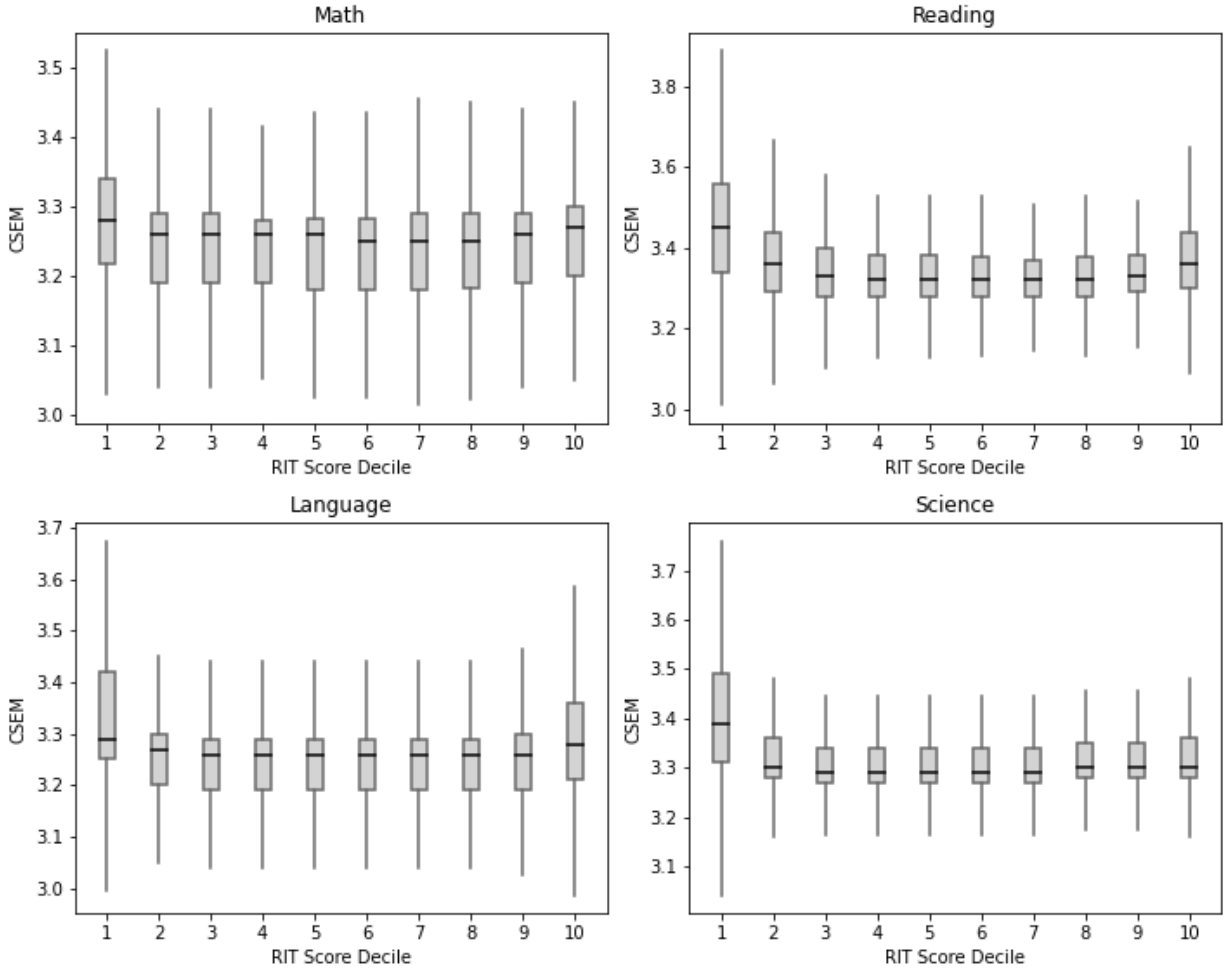


Table 7.5. CSEM Summary by Subject, Grade, and Decile Score Group

Subject	Grade	Fall Score Groups			Winter Score Groups			Spring Score Groups		
		1st	5th	10th	1st	5th	10th	1st	5th	10th
Math	K	3.36	3.26	3.36	3.36	3.27	3.29	3.36	3.27	3.27
Math	1	3.31	3.25	3.31	3.32	3.26	3.29	3.33	3.26	3.29
Math	2	3.32	3.23	3.30	3.33	3.24	3.28	3.35	3.24	3.29
Math	3	3.30	3.22	3.26	3.30	3.24	3.25	3.31	3.24	3.25
Math	4	3.29	3.22	3.26	3.27	3.23	3.26	3.28	3.24	3.26
Math	5	3.28	3.23	3.26	3.27	3.24	3.26	3.27	3.24	3.26
Math	6	3.32	3.23	3.26	3.32	3.25	3.28	3.34	3.24	3.27
Math	7	3.34	3.24	3.27	3.31	3.25	3.29	3.33	3.25	3.28
Math	8	3.34	3.24	3.28	3.32	3.25	3.32	3.33	3.25	3.32
Math	9	3.39	3.27	3.38	3.35	3.28	3.43	3.37	3.28	3.43
Math	10	3.38	3.27	3.40	3.36	3.29	3.48	3.37	3.28	3.46
Math	11	3.39	3.27	3.42	3.37	3.29	3.49	3.39	3.28	3.48
Math	12	3.44	3.28	3.41	3.42	3.29	3.52	3.43	3.28	3.53

Subject	Grade	Fall Score Groups			Winter Score Groups			Spring Score Groups		
		1st	5th	10th	1st	5th	10th	1st	5th	10th
Reading	K	3.30	3.26	3.37	3.30	3.25	3.30	3.30	3.25	3.29
Reading	1	3.29	3.26	3.30	3.28	3.26	3.30	3.29	3.26	3.31
Reading	2	3.53	3.31	3.42	3.59	3.32	3.40	3.60	3.31	3.42
Reading	3	3.56	3.35	3.38	3.53	3.36	3.38	3.56	3.36	3.37
Reading	4	3.50	3.34	3.39	3.47	3.35	3.40	3.48	3.35	3.42
Reading	5	3.48	3.34	3.43	3.46	3.35	3.48	3.46	3.35	3.53
Reading	6	3.54	3.34	3.36	3.53	3.34	3.36	3.54	3.34	3.38
Reading	7	3.55	3.34	3.38	3.53	3.34	3.38	3.53	3.34	3.39
Reading	8	3.53	3.34	3.40	3.51	3.34	3.42	3.52	3.34	3.44
Reading	9	3.59	3.34	3.44	3.55	3.35	3.46	3.56	3.35	3.47
Reading	10	3.57	3.34	3.49	3.54	3.35	3.51	3.55	3.35	3.51
Reading	11	3.59	3.35	3.52	3.56	3.36	3.54	3.57	3.36	3.53
Reading	12	3.61	3.37	3.53	3.59	3.37	3.56	3.60	3.38	3.55
Language	2	3.47	3.26	3.35	3.59	3.26	3.27	3.71	3.27	3.28
Language	3	3.40	3.25	3.29	3.40	3.25	3.26	3.44	3.26	3.27
Language	4	3.36	3.24	3.27	3.35	3.24	3.26	3.36	3.26	3.28
Language	5	3.35	3.24	3.27	3.33	3.25	3.26	3.34	3.25	3.29
Language	6	3.35	3.25	3.27	3.33	3.24	3.28	3.34	3.26	3.30
Language	7	3.34	3.25	3.29	3.32	3.25	3.31	3.33	3.27	3.33
Language	8	3.33	3.25	3.36	3.31	3.25	3.40	3.33	3.27	3.43
Language	9	3.40	3.26	3.39	3.35	3.26	3.45	3.37	3.28	3.48
Language	10	3.37	3.26	3.51	3.35	3.26	3.58	3.37	3.29	3.61
Language	11	3.38	3.27	3.62	3.35	3.27	3.68	3.39	3.29	3.67
Language	12	3.37	3.29	3.68	3.37	3.28	3.73	3.38	3.29	3.70
Science	2	3.57	3.34	3.40	3.82	3.33	3.34	3.76	3.29	3.29
Science	3	3.47	3.32	3.37	3.53	3.33	3.33	3.50	3.31	3.30
Science	4	3.44	3.32	3.33	3.45	3.33	3.34	3.41	3.30	3.31
Science	5	3.43	3.32	3.35	3.43	3.32	3.35	3.40	3.30	3.32
Science	6	3.43	3.31	3.35	3.44	3.31	3.35	3.44	3.30	3.31
Science	7	3.45	3.31	3.34	3.45	3.32	3.35	3.43	3.31	3.32
Science	8	3.45	3.32	3.35	3.44	3.32	3.37	3.42	3.31	3.35
Science	9	3.43	3.31	3.34	3.43	3.30	3.36	3.46	3.32	3.38
Science	10	3.47	3.31	3.35	3.45	3.30	3.36	3.47	3.32	3.40
Science	11	3.49	3.31	3.36	3.50	3.31	3.39	3.53	3.33	3.43
Science	12	3.56	3.31	3.38	3.60	3.31	3.40	3.62	3.32	3.44

Note. Language = Language Usage

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

Subject	Demographic Group	Fall Score Groups			Winter Score Groups			Spring Score Groups		
		1st	5th	10th	1st	5th	10th	1st	5th	10th
Math	Female	3.31	3.24	3.31	3.30	3.25	3.31	3.31	3.25	3.29
Math	Male	3.33	3.24	3.29	3.32	3.25	3.29	3.34	3.25	3.28
Math	White	3.32	3.24	3.28	3.32	3.25	3.28	3.34	3.25	3.27
Math	Black	3.32	3.24	3.41	3.31	3.25	3.40	3.32	3.25	3.35
Math	Asian	3.35	3.24	3.29	3.35	3.25	3.29	3.36	3.25	3.28
Math	Hispanic	3.32	3.24	3.34	3.31	3.25	3.35	3.32	3.25	3.31
Math	AI/AN	3.31	3.25	3.37	3.31	3.25	3.37	3.31	3.25	3.32
Math	NH/PI	3.33	3.25	3.33	3.33	3.26	3.36	3.34	3.26	3.31
Math	Multiple	3.32	3.24	3.30	3.31	3.25	3.30	3.32	3.25	3.29
Math	Other	3.33	3.24	3.32	3.32	3.25	3.30	3.34	3.25	3.30
Reading	Female	3.49	3.33	3.41	3.49	3.33	3.41	3.50	3.33	3.41
Reading	Male	3.50	3.33	3.41	3.50	3.33	3.40	3.52	3.34	3.41
Reading	White	3.48	3.33	3.40	3.49	3.33	3.40	3.51	3.33	3.40
Reading	Black	3.49	3.34	3.44	3.48	3.34	3.44	3.50	3.33	3.42
Reading	Asian	3.49	3.33	3.40	3.50	3.33	3.40	3.51	3.33	3.41
Reading	Hispanic	3.51	3.33	3.41	3.51	3.33	3.41	3.52	3.34	3.40
Reading	AI/AN	3.49	3.33	3.43	3.48	3.33	3.40	3.49	3.33	3.40
Reading	NH/PI	3.50	3.33	3.41	3.50	3.33	3.40	3.51	3.33	3.41
Reading	Multiple	3.48	3.33	3.40	3.49	3.33	3.40	3.50	3.34	3.41
Reading	Other	3.49	3.32	3.43	3.50	3.32	3.42	3.51	3.33	3.41
Language	Female	3.37	3.25	3.35	3.36	3.25	3.34	3.37	3.26	3.35
Language	Male	3.37	3.25	3.35	3.36	3.26	3.34	3.37	3.27	3.35
Language	White	3.36	3.25	3.34	3.35	3.25	3.33	3.37	3.27	3.34
Language	Black	3.36	3.24	3.34	3.35	3.24	3.35	3.36	3.26	3.36
Language	Asian	3.38	3.25	3.35	3.37	3.25	3.35	3.38	3.26	3.35
Language	Hispanic	3.40	3.25	3.37	3.38	3.25	3.35	3.39	3.26	3.36
Language	AI/AN	3.36	3.26	3.43	3.35	3.26	3.38	3.36	3.27	3.38
Language	NH/PI	3.37	3.25	3.39	3.37	3.26	3.36	3.38	3.27	3.36
Language	Multiple	3.35	3.25	3.36	3.35	3.25	3.34	3.36	3.26	3.35
Language	Other	3.36	3.25	3.36	3.35	3.25	3.35	3.37	3.26	3.36
Science	Female	3.44	3.32	3.35	3.45	3.32	3.36	3.43	3.31	3.33
Science	Male	3.46	3.32	3.35	3.46	3.32	3.35	3.45	3.31	3.32
Science	White	3.44	3.31	3.34	3.44	3.31	3.34	3.43	3.31	3.32
Science	Black	3.44	3.32	3.38	3.44	3.33	3.40	3.43	3.31	3.36
Science	Asian	3.49	3.32	3.35	3.51	3.32	3.35	3.46	3.30	3.31
Science	Hispanic	3.46	3.32	3.35	3.47	3.32	3.37	3.44	3.30	3.33
Science	AI/AN	3.44	3.31	3.35	3.45	3.32	3.37	3.44	3.31	3.34
Science	NH/PI	3.45	3.32	3.35	3.48	3.32	3.35	3.44	3.31	3.35
Science	Multiple	3.45	3.32	3.36	3.45	3.32	3.36	3.44	3.30	3.33
Science	Other	3.43	3.31	3.34	3.42	3.31	3.33	3.44	3.33	3.35

Note. Language = Language Usage

7.6. Score Reliability

The CSEM is dependent on the underlying scale. If the scale is increased by multiplying scores by 10, for example, then the CSEM values will also increase because they too should be multiplied by 10. Score reliability overcomes this limitation of the CSEM 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. As such, reliability may be used to compare the precision of scores obtained from different assessments.

Test information and CSEM values are computed for individual students. 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 close to 1 indicate no measurement error and highly consistent scores. Conversely, values close to zero indicate either no true score variance or a substantial amount of error variance. There are multiple types of reliability estimates, and each type of estimate reflects the influence of a particular source of error on the consistency of scores. A *test-retest reliability* estimate is a coefficient of stability. It captures the influence of time on score consistency. Test-retest is calculated as the correlation of scores from one occasion with scores from the same test obtained from the same students obtained on a second occasion. A variation of this estimate is test-retest with alternate forms, where the tests given on each occasion are very similar but not identical test forms. This type of reliability estimate is impacted by 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 a common way to estimate internal consistency for a group of students. It is calculated as shown in Equation 9.

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

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 a large amount of true score variance (e.g., all fourth-grade students in the United States) will have higher reliability estimates than a homogenous group of students with a low amount of true score variance (e.g., first-grade students with dyslexia). Because the reliability coefficient depends on the group of examinees, it is necessary to estimate reliability for all major subgroups taking a test, such as groups representing different races and ethnicities. Reliability estimates should be high and of similar magnitude for all groups of students.

7.6.1. Marginal Reliability Results

Marginal reliability estimates for MAP Growth were examined across several factors, including subject, instructional area, term (fall, winter, spring), grade, and demographic. Results from the 2024–2025 administration indicate that MAP Growth provides consistently strong score

reliability across the studied factors, with reliability estimates generally meeting or exceeding commonly accepted thresholds for educational decision-making.

At the subject level, marginal reliabilities were uniformly high across terms, as shown in Table 7.7. Math and Reading reliabilities were generally mid-to-high 0.90 across most grades, with lower values in kindergarten (e.g., Math: 0.92–0.94; Reading: 0.90–0.94) and higher values in upper grades (often 0.97–0.98). Language Usage reliabilities were especially stable, at 0.96 for grades 2–8 and 0.97 for grades 9–12 across terms. Science reliabilities were slightly lower overall, generally 0.93–0.96 across grades and terms.

Across terms, marginal reliability was generally stable, with a slight tendency for spring administrations to produce marginally higher reliability estimates than fall, particularly in Math. This pattern is consistent with broader ability distributions and increased test engagement later in the academic year. Differences across terms, however, were small and do not meaningfully affect score interpretation.

Instructional area names differ from state to state, even when the content of the instructional area is the same. Calculating marginal reliability for each instructional area name leads to an extensive set of results that spans hundreds of pages. While the full results were examined, only a summary is included herein. As shown in Table 7.8, across instructional areas, the average marginal reliability was between 0.86 and 0.91. Even the lowest observed marginal reliability estimates were 0.75 or higher. The lower reliability estimates were seen for instructional areas with small numbers of items. Examination of the detailed results file indicates that instructional areas with larger numbers of items and broader construct coverage (e.g., Operations and Algebraic Thinking, Geometry, Informational Text, Literary Text, Vocabulary, and comprehensive Language/Writing strands) consistently yield higher marginal reliability, frequently in the low-to-mid 0.90s across states and terms. These instructional areas benefit from greater measurement precision due to having more items and wider ability coverage.

Overall, the findings demonstrate that MAP Growth instructional area scores provide reliable measurement across subjects, grades, and terms. While variability exists at the most granular instructional-area level—particularly for narrowly defined constructs with limited data—these instances are well understood from a psychometric perspective and do not detract from the overall robustness of MAP Growth scores. The consistency of results across the comprehensive instructional-area dataset reinforces the validity of reporting instructional area scores for instructional planning and research purposes.

MAP Growth marginal reliability estimates were consistently high within each demographic subgroup, with only very small differences by season, as shown in Table 7.9. In Math, marginal reliability was 0.99 for every reported subgroup (female, male, and each race/ethnicity category) in fall, winter, and spring, indicating uniformly strong score precision across examinee demographic groups and testing windows. Note that reliability is higher in the demographic subgroup analysis because scores are combined across grades, which leads to more true-score variance than an analysis that separates students by grade.

In Reading, marginal reliability was similarly strong, with values of 0.99 in fall for nearly all groups and 0.98 in winter and spring for those same groups; the NH/PI subgroup showed 0.98 across fall, winter, and spring. Language Usage marginal reliability was slightly lower but still high, ranging from 0.96 to 0.98 across groups and seasons (e.g., male, Black, Hispanic, and AI/AN groups reached 0.98 in fall, while the White group was 0.96 in spring). Science marginal reliability was also high and stable, ranging from 0.95 to 0.96 across demographic subgroups

and seasons, with most groups at 0.95 and some at 0.96 depending on season (e.g., males at 0.96 across seasons). Overall, these results support that MAP Growth score precision is consistently strong across demographic subgroups and is comparable across fall, winter, and spring administrations within each subject.

As shown in Appendix A, marginal reliability by state and term was in the upper 0.90s. Note that results for a state were omitted when the sample size was below 5,000 for any term.

Table 7.7. Marginal Reliability by Subject, Grade, and Term

Subject	Grade	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Math	K	613,720	0.92	656,491	0.93	695,070	0.94
Math	1	824,640	0.94	795,619	0.94	833,281	0.95
Math	2	934,163	0.95	902,976	0.95	936,139	0.95
Math	3	997,802	0.95	951,035	0.95	936,819	0.96
Math	4	974,054	0.96	922,787	0.96	904,212	0.96
Math	5	983,489	0.96	932,039	0.96	910,198	0.97
Math	6	1,006,521	0.96	911,662	0.96	909,925	0.97
Math	7	1,003,779	0.97	887,365	0.97	892,846	0.97
Math	8	905,110	0.97	794,870	0.97	775,508	0.98
Math	9	390,393	0.97	298,018	0.98	317,446	0.98
Math	10	311,224	0.98	239,780	0.98	256,758	0.98
Math	11	193,049	0.98	147,097	0.98	139,921	0.98
Math	12	97,571	0.98	69,176	0.98	53,125	0.98
Reading	K	483,763	0.90	531,217	0.93	568,122	0.94
Reading	1	681,966	0.95	661,049	0.95	694,311	0.96
Reading	2	821,598	0.96	797,782	0.96	840,132	0.96
Reading	3	969,558	0.96	923,437	0.96	914,423	0.96
Reading	4	964,004	0.96	902,142	0.96	884,051	0.96
Reading	5	976,102	0.96	916,843	0.96	897,520	0.96
Reading	6	1,043,354	0.96	943,338	0.96	932,885	0.96
Reading	7	1,048,553	0.96	926,162	0.96	915,552	0.96
Reading	8	1,033,878	0.96	911,629	0.96	882,519	0.96
Reading	9	679,174	0.96	543,312	0.96	548,215	0.96
Reading	10	560,670	0.96	440,154	0.96	448,997	0.96
Reading	11	308,760	0.96	229,435	0.97	211,410	0.97
Reading	12	169,964	0.97	117,606	0.97	95,899	0.97
Language	2	105,876	0.96	95,040	0.96	109,126	0.96
Language	3	179,672	0.96	155,061	0.96	169,899	0.96
Language	4	181,775	0.96	157,674	0.96	170,140	0.96
Language	5	188,845	0.96	163,244	0.96	178,211	0.96
Language	6	204,596	0.96	166,770	0.96	187,079	0.96
Language	7	199,913	0.96	159,574	0.96	181,393	0.96
Language	8	196,781	0.96	156,378	0.96	173,762	0.96
Language	9	113,841	0.97	74,537	0.97	90,116	0.97

Subject	Grade	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Language	10	98,067	0.97	65,649	0.97	78,597	0.97
Language	11	66,557	0.97	43,854	0.97	46,129	0.97
Language	12	37,387	0.97	24,437	0.97	20,505	0.97
Science	2	37,898	0.94	36,353	0.93	39,377	0.94
Science	3	159,088	0.93	135,128	0.93	157,750	0.94
Science	4	218,933	0.93	181,424	0.93	220,759	0.94
Science	5	318,298	0.94	280,192	0.94	300,349	0.94
Science	6	288,543	0.94	242,384	0.94	275,674	0.95
Science	7	304,969	0.94	258,698	0.95	294,632	0.95
Science	8	336,583	0.94	287,273	0.95	301,529	0.95
Science	9	60,703	0.94	47,127	0.95	52,234	0.95
Science	10	56,583	0.94	45,144	0.95	45,862	0.96
Science	11	39,081	0.95	30,375	0.95	26,323	0.96
Science	12	15,647	0.95	11,402	0.96	9,089	0.96

Note. Language = Language Usage

Table 7.8. Marginal Reliability of Instructional Area Scores by Subject and Term

Subject	Term	Marginal Reliability Summary		
		Minimum	Mean	Maximum
Math	Fall	0.75	0.91	0.97
Math	Winter	0.76	0.91	0.96
Math	Spring	0.80	0.91	0.96
Reading	Fall	0.80	0.88	0.96
Reading	Winter	0.79	0.88	0.95
Reading	Spring	0.78	0.87	0.95
Language	Fall	0.85	0.91	0.96
Language	Winter	0.83	0.91	0.96
Language	Spring	0.78	0.89	0.96
Science	Fall	0.78	0.86	0.90
Science	Winter	0.77	0.86	0.91
Science	Spring	0.81	0.86	0.92

Note. Language = Language Usage

Table 7.9. Marginal Reliability by Subject, Demographics Group, and Term

Subject	Demographic Group	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Math	Female	4,538,716	0.99	4,180,535	0.99	4,211,701	0.99
Math	Male	4,693,379	0.99	4,326,331	0.99	4,346,932	0.99
Math	White	3,848,470	0.99	3,525,287	0.99	3,530,786	0.99
Math	Black	1,376,021	0.99	1,307,584	0.99	1,285,100	0.99
Math	Asian	482,150	0.99	447,427	0.99	458,013	0.99
Math	Hispanic	2,315,989	0.99	2,186,998	0.99	2,156,491	0.99
Math	AI/AN	131,418	0.99	120,316	0.99	111,557	0.99
Math	NH/PI	25,674	0.99	23,243	0.99	22,093	0.99
Math	Multiple	464,080	0.99	427,789	0.99	456,426	0.99
Math	Other	588,293	0.99	468,222	0.99	538,167	0.99
Reading	Female	4,768,794	0.99	4,326,824	0.98	4,330,425	0.98
Reading	Male	4,968,452	0.99	4,514,553	0.98	4,500,464	0.98
Reading	White	4,026,300	0.99	3,643,085	0.98	3,618,180	0.98
Reading	Black	1,507,577	0.99	1,408,470	0.98	1,362,134	0.98
Reading	Asian	516,365	0.99	472,416	0.98	476,660	0.98
Reading	Hispanic	2,440,333	0.99	2,250,029	0.98	2,221,516	0.98
Reading	AI/AN	131,941	0.99	118,451	0.98	110,335	0.98
Reading	NH/PI	27,963	0.98	25,264	0.98	23,672	0.98
Reading	Multiple	480,836	0.99	438,149	0.98	465,450	0.98
Reading	Other	605,931	0.99	485,513	0.98	552,942	0.98
Language	Female	774,579	0.97	621,470	0.97	692,779	0.97
Language	Male	798,006	0.98	640,582	0.97	711,482	0.97
Language	White	717,026	0.97	569,429	0.97	636,044	0.96
Language	Black	194,033	0.98	172,946	0.97	183,646	0.97
Language	Asian	58,216	0.97	52,184	0.97	59,757	0.97
Language	Hispanic	303,746	0.98	240,610	0.98	251,375	0.97
Language	AI/AN	37,918	0.98	33,802	0.97	29,503	0.97
Language	NH/PI	3,649	0.97	2,968	0.97	3,134	0.97
Language	Multiple	76,183	0.97	59,757	0.97	71,014	0.97

Subject	Demographic Group	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Language	Other	181,814	0.97	130,356	0.97	169,788	0.97
Science	Female	903,327	0.95	763,722	0.95	848,366	0.95
Science	Male	932,856	0.96	791,671	0.96	875,118	0.96
Science	White	680,767	0.95	570,393	0.95	623,326	0.95
Science	Black	281,083	0.95	246,624	0.95	267,327	0.96
Science	Asian	93,330	0.96	80,809	0.96	91,852	0.96
Science	Hispanic	558,027	0.96	477,317	0.95	525,144	0.96
Science	AI/AN	39,377	0.95	33,251	0.95	33,548	0.96
Science	NH/PI	4,734	0.96	4,013	0.96	4,274	0.96
Science	Multiple	86,200	0.95	72,188	0.95	87,636	0.96
Science	Other	92,665	0.95	70,798	0.95	90,377	0.95

Note. Language = Language Usage

7.6.2. Test-Retest Reliability Results

MAP Growth affords the means to assess students on multiple occasions (e.g., fall, winter, and spring) during the school year. Thus, test-retest reliability is key, as it provides insight into the consistency of MAP Growth across time. The adaptive nature of MAP Growth assessments requires reliability to be examined using non-traditional methods because dynamic item selection is an integral part of MAP Growth. Parallel forms are restricted to identical item content from a common instructional area, but the item difficulties depend on the student's responses to previous items on the test. Therefore, test-retest reliability of MAP Growth is more accurately described as a mix between test-retest reliability and a type of alternate forms reliability, both of which are spread across several months versus the typical two or three weeks. The second test (or retest) is not the same test; rather, it is comparable to the first by its content and structure, differing only in the difficulty level of its items. In other words, test-retest with alternate forms (Crocker & Algina, 1986) describes the influence of two sources of measurement error: time and item selection.

Test-retest with alternate forms reliability for MAP Growth was estimated via the Pearson correlation between MAP Growth RIT scores of students taking MAP Growth in two consecutive terms (e.g., fall and winter, fall and spring, winter and spring).

Test-retest reliability analyses were conducted to evaluate the temporal stability of MAP Growth scores across testing windows within the same academic year. Reliability coefficients were examined for fall–winter, fall–spring, and winter–spring 2024–2025 administrations by subject and grade, as well as across key demographic subgroups. The results are shown in Table 7.10 and Table 7.11.

At the grade level, MAP Growth demonstrated moderate to strong test-retest reliability across subjects and grade bands. Reliability was generally highest for winter–spring comparisons, followed by fall–winter, with fall–spring comparisons yielding the lowest coefficients. This pattern is consistent with expectations, as fall–spring comparisons span a longer interval during which true academic growth is expected to occur, thereby attenuating stability coefficients.

In Math, test-retest reliability increased steadily from early elementary through the middle grades, peaking in grades 4–8, where coefficients typically ranged from the high 0.80s to low 0.90s depending on the comparison window. Reliability declined modestly in high school grades, particularly for fall–spring comparisons, reflecting both increased instructional differentiation and longer intervals between measurements. Kindergarten Math exhibited lower reliability than later grades, a pattern consistent with greater developmental variability and emerging skill acquisition at early ages.

Reading followed a similar trajectory, with reliability improving rapidly from kindergarten into elementary grades and remaining strong through middle school. As with Math, fall–spring reliability was consistently lower than fall–winter and winter–spring estimates, particularly in secondary grades. Early-grade Reading (especially kindergarten) showed the lowest stability coefficients, reflecting rapid skill development and increased measurement error associated with emergent literacy skills.

For Language Usage, test-retest reliability estimates were generally strong across grades 2–8 and remained acceptable through high school. Patterns across comparison windows mirrored those observed in Math and Reading, with shorter intervals yielding higher stability. Reliability

declined gradually in grades 9–12, consistent with increasing curricular specialization and variation in instructional exposure.

Science test-retest reliability was slightly lower on average than for Math, Reading, and Language Usage but remained within acceptable ranges for most grades. Reliability was strongest in upper elementary and middle grades and declined in high school, particularly for fall–spring comparisons. As with other subjects, winter–spring reliability tended to exceed fall–spring estimates, reflecting the shorter interval between administrations.

Analyses by demographic subgroup indicate that MAP Growth scores exhibit high and consistent temporal stability across student groups. Across Math and Reading, test-retest reliability coefficients were uniformly high (generally above 0.90 for shorter intervals) for all reported gender and race/ethnicity groups, with only minor differences observed among subgroups. Language Usage and Science reliability estimates were slightly lower overall but remained comparable across demographic categories. Importantly, no subgroup demonstrated systematically weaker test-retest reliability, supporting the conclusion that MAP Growth scores function consistently across diverse populations.

Overall, the test-retest results provide strong evidence that MAP Growth scores are stable over time when true growth is expected to be limited and behave as anticipated when longer intervals allow for instructional effects and learning. These findings support the use of MAP Growth for monitoring student progress across seasons and for longitudinal analyses of academic growth at both the individual and group levels.

Table 7.10. Test-Retest Reliability by Subject, Grade, and Term

Subject	Grade	N	Fall/Winter	Fall/Spring	Winter/Spring
			Reliability	Reliability	Reliability
Math	K	529,090	0.76	0.70	0.80
Math	1	715,735	0.83	0.79	0.86
Math	2	806,356	0.86	0.82	0.88
Math	3	813,405	0.88	0.85	0.89
Math	4	785,629	0.90	0.87	0.90
Math	5	788,428	0.91	0.89	0.92
Math	6	740,051	0.90	0.87	0.91
Math	7	707,251	0.90	0.88	0.90
Math	8	600,893	0.89	0.85	0.88
Math	9	184,366	0.84	0.76	0.81
Math	10	154,672	0.82	0.74	0.80
Math	11	82,731	0.82	0.74	0.79
Math	12	33,132	0.82	0.72	0.78
Reading	K	411,722	0.71	0.65	0.77
Reading	1	585,458	0.84	0.80	0.87
Reading	2	704,803	0.86	0.82	0.88
Reading	3	779,997	0.87	0.84	0.88
Reading	4	765,670	0.88	0.86	0.89
Reading	5	771,549	0.89	0.87	0.89
Reading	6	758,125	0.88	0.86	0.87

Subject	Grade	N	Fall/Winter	Fall/Spring	Winter/Spring
			Reliability	Reliability	Reliability
Reading	7	723,539	0.87	0.84	0.86
Reading	8	695,845	0.85	0.82	0.84
Reading	9	338,529	0.82	0.74	0.79
Reading	10	280,966	0.80	0.72	0.76
Reading	11	123,927	0.78	0.69	0.74
Reading	12	56,136	0.76	0.67	0.73
Language	2	80,338	0.86	0.82	0.88
Language	3	123,323	0.88	0.86	0.89
Language	4	124,050	0.88	0.87	0.89
Language	5	128,822	0.89	0.87	0.89
Language	6	127,305	0.88	0.87	0.88
Language	7	119,330	0.88	0.86	0.87
Language	8	114,401	0.87	0.84	0.85
Language	9	48,197	0.84	0.79	0.82
Language	10	43,370	0.83	0.77	0.80
Language	11	25,674	0.82	0.74	0.78
Language	12	13,038	0.78	0.70	0.76
Science	2	32,158	0.80	0.75	0.83
Science	3	115,904	0.84	0.81	0.85
Science	4	152,636	0.84	0.82	0.86
Science	5	229,237	0.85	0.82	0.85
Science	6	199,515	0.85	0.83	0.85
Science	7	207,582	0.85	0.82	0.84
Science	8	220,530	0.82	0.77	0.79
Science	9	27,372	0.79	0.72	0.76
Science	10	26,775	0.78	0.70	0.74
Science	11	16,468	0.75	0.66	0.72
Science	12	5,889	0.74	0.68	0.74

Note. Language = Language Usage

Table 7.11. Test-Retest Reliability by Subject, Demographic Group, and Term

Subject	Demographic Group	N	Fall/Winter	Fall/Spring	Winter/Spring
			Reliability	Reliability	Reliability
Math	Female	3,426,448	0.96	0.94	0.95
Math	Male	3,527,427	0.96	0.93	0.95
Math	Asian	367,681	0.97	0.95	0.97
Math	Black	1,022,250	0.95	0.91	0.93
Math	NH/PI	16,530	0.96	0.93	0.95
Math	AI/AN	87,072	0.95	0.92	0.94
Math	Hispanic	1,735,062	0.95	0.92	0.94
Math	Other	347,902	0.96	0.94	0.96
Math	Multiple	336,040	0.96	0.94	0.95
Math	White	2,891,324	0.96	0.94	0.96

Subject	Demographic Group	N	Fall/Winter	Fall/Spring	Winter/Spring
			Reliability	Reliability	Reliability
Reading	Female	3,441,780	0.95	0.93	0.94
Reading	Male	3,569,800	0.94	0.92	0.93
Reading	Asian	377,236	0.96	0.94	0.95
Reading	Black	1,058,634	0.93	0.91	0.92
Reading	NH/PI	17,098	0.95	0.92	0.94
Reading	AI/AN	84,076	0.94	0.92	0.93
Reading	Hispanic	1,726,841	0.94	0.91	0.93
Reading	Other	344,726	0.95	0.93	0.94
Reading	Multiple	334,942	0.95	0.93	0.94
Reading	White	2,922,947	0.95	0.93	0.94
Language	Female	470,277	0.92	0.89	0.91
Language	Male	480,117	0.91	0.87	0.90
Language	Asian	36,467	0.93	0.91	0.92
Language	Black	125,637	0.90	0.87	0.89
Language	NH/PI	1,816	0.92	0.90	0.92
Language	AI/AN	20,538	0.90	0.87	0.89
Language	Hispanic	171,820	0.90	0.87	0.89
Language	Other	97,166	0.92	0.89	0.91
Language	Multiple	41,787	0.91	0.88	0.90
Language	White	425,678	0.91	0.89	0.90
Science	Female	608,467	0.87	0.83	0.86
Science	Male	628,441	0.88	0.84	0.87
Science	Asian	67,511	0.90	0.88	0.90
Science	Black	188,398	0.84	0.80	0.83
Science	NH/PI	2,887	0.88	0.83	0.85
Science	AI/AN	23,505	0.86	0.82	0.84
Science	Hispanic	369,952	0.86	0.82	0.85
Science	Other	51,596	0.88	0.85	0.87
Science	Multiple	51,613	0.87	0.84	0.86
Science	White	438,512	0.87	0.84	0.86

Note. Language = Language Usage

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

A scale score is a transformation of the examinee ability parameter, θ. Scale scores alone are just a set of numbers. Additional information or transformation is needed to facilitate the interpretation of scores. *Norm-referenced* information is provided by comparing a student's performance with the performance 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. Tests 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 scores are reported on the RIT (**R**asch un**IT**) scale. Scores have a mean of 200, a standard deviation 10, and a typical range from 100 to 350. A momentary RIT score is calculated after an examinee responds to a test question. The momentary score is used 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 a preliminary student score based on past test performance. If a student has no prior test score, a default starting value is assigned according to test subject and the student's grade. As each 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. Momentary ability estimates are updated after each response using Bayesian methods (Owen, 1975) that consider all of the student's responses up to that point in the test. The updated momentary ability estimate is factored into the selection of the next item. As this cycle is repeated, each successive momentary ability estimate is slightly more precise than the previous one. The test continues until the minimum test length is reached and standard error associated with the estimate is within acceptable limits or 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 calculated on the logit metric and then converted to a RIT score by linear transformation.

A limitation of item response theory is that a maximum likelihood estimate of an examinee's ability is not estimable when the examinee answers every item correctly or every item incorrectly. Maximum likelihood with fencing (MLEF) overcomes this limitation by using two fictional (fence) items. The lower-fence item is selected to have a difficulty 3.8 logits below the easiest item answered by the examinee, and a correct response is assigned to it. The upper-fence item is chosen to have a difficulty value 3.8 logits above the most difficult item completed by the examinee, and an incorrect response is assigned to it. These fixed item parameters are used to calculate the probability of a response for the lower-fence item, $P_{LF}(\theta)$, and the probability of a response to the upper-fence item, $P_{UF}(\theta)$. The MLEF estimate is obtained by finding the value of theta that maximizes the log-likelihood of the item response pattern, **x**, and the fencing items, as shown in Equation 10.

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

This can be done using numerical optimization (e.g., Newton-Raphson). The fenced estimator of a student's score is calculated as shown in Equation 11.

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

This is linearly transformed to the RIT scale using $RIT = 10\hat{\theta}_{MLEF} + 200$.

8.1.3. Instructional Area Scores

An overall RIT score is calculated using an examinee's entire item response pattern. Instructional area scores are calculated using only responses to items for a specific instructional area. The instructional area score is calculated with MLEF in the same way as the overall RIT, but the response pattern is limited to a subset of the items. Instructional area scores are converted to the RIT scale using the same linear transformation as the overall RIT score.

8.2. MAP Growth Norms

Meaningful interpretation of MAP Growth scores requires appropriate contextualization through comparison with established references. Without this context, assessment scores remain isolated numbers rather than actionable information. MAP Growth Achievement and Growth norms serve a descriptive function, characterizing the achievement and growth patterns of U.S. public school students to provide interpretive context for individual and group performance. The 2025 MAP Growth norms (NWEA, 2025) provide a norm-referenced interpretation of student achievement and growth.

NWEA continuously refines the statistical models underlying MAP Growth achievement and growth norms to enhance their accuracy and validity. The 2025 norms update addresses several critical factors: evolving U.S. student demographics, documented shifts in academic performance following the COVID-19 pandemic, and improvements in MAP Growth's Enhanced Item Selection Algorithm (EISA). The updated norms enable educators to (1) evaluate student and school achievement and growth, (2) differentiate instruction and set goals with students, and (3) support conversations about achievement and growth patterns in light of these critical factors.

Data used to produce the 2025 MAP Growth national norms were sampled from 116 million scores of 13.8 million students across 30,000 schools spanning 6 testing terms from fall 2022 to spring 2024. The sample used broadly reflects the U.S. population. In total, 344 growth models were evaluated to identify 86 unique models for both student and school achievement and growth norms. The *2025 Norms Quick Reference* guide provides a succinct overview of the norms. A complete description of the 2025 norms is provided in the full *2025 MAP Growth Norms Technical Manual* (NWEA, 2025).

8.3. Linking Studies and Predicted Proficiency

State summative assessments have multiple performance levels (e.g., *Proficient*, *Advanced*) that are established by a state through a formal standard setting process. Scores that divide a scale into these levels are referred to as cut scores.

Because most MAP Growth tests are aligned to state standards, student RIT scores are strongly related to performance on state summative assessments. Linking studies use this relationship to connect MAP Growth scores to state test performance levels.

Most linking studies are conducted in grades 3–8 for English language arts (ELA) and mathematics, and sometimes for high school or science when sufficient data are available. Since most states do not test grade 2 students, grade 2 MAP Growth scores are linked to the grade 3 state assessment.

Linking studies allow educators to use MAP Growth results from fall, winter, and early spring to estimate how students are likely to perform on state summative tests administered later in the spring, or in the case of grade 2 students, in the following year. This information helps schools identify students who may be at risk of not meeting proficiency and act through targeted instruction or intervention prior to the state summative assessment.

In states with “Read by Grade 3” policies, linking results—especially for grade 2 Reading—can help identify students who need additional support. In cases where states use specific grade 3 Reading cut scores for decisions such as retention, additional linking studies can be conducted to identify corresponding RIT scores that help guide instructional and placement decisions.

8.3.1. Linking Study Methodology

Linking studies use student data from the same spring administration of the MAP Growth and state summative assessments. Only students who took both assessments during that spring term are included. To ensure reliable results, at least 1,000 students per grade and subject are required. Participation is voluntary, and districts must provide state test data and permission for the use of MAP Growth results. Student records are matched using identifying information such as name and student ID.

Because not all students in a state take MAP Growth, the study sample may not represent the overall state population. To address this, statistical weighting is applied so that the sample reflects the state’s student population in terms of race, gender, and performance level. These factors are closely related to academic outcomes and are commonly reported by states. This adjustment allows the results of the study to be used confidently for students statewide.

To establish the link between MAP Growth scores and state test performance levels, an equipercentile linking method is used. This approach matches scores from the two assessments that represent the same percentile of students. In other words, a MAP Growth score is identified that corresponds to the same proportion of students meeting or exceeding a given cut score on the state test. This process produces spring MAP Growth cut scores on the RIT scale that correspond to cut scores on the state assessment scale. As a result, students may be classified into state performance levels using MAP Growth scores.

Once spring cut scores are established, growth norms are used to project corresponding cut scores for fall and winter. These norms describe typical student growth between testing seasons, allowing educators to estimate whether students are on track to meet spring proficiency targets earlier in the year.

Since most states do not test students in grade 2, grade 2 cut scores are developed using data from the same group of students as they move from grade 2 into grade 3. Growth patterns from

this cohort are used to estimate grade 2 fall, winter, and spring MAP Growth cut scores that align with grade 3 state expectations.

Overall, this methodology allows schools to use MAP Growth results throughout the year to predict state test outcomes, identify students who may need additional support, and make informed instructional decisions well before spring testing.

8.3.2. Linking Study Accuracy in Predicting Proficiency

The degree to which MAP Growth predicts student proficiency status on the state summative tests can be described using classification accuracy statistics based on the MAP Growth spring RIT cut scores obtained by linking to the state cut scores. The results show the proportion of students correctly classified by their RIT scores as proficient or not proficient on the state summative test. A summary of how well the interpolated grade 2 cuts predict grade 3 proficiency status is also reported in the classification accuracy statistics.

Detailed linking study reports for each participating state are available on the NWEA Connection website. Each report provides details of the methodology and student sample, as well as comprehensive details on the classification accuracy analysis, including the false positive rate, false negative rate, sensitivity, specificity, precision, and area under the curve statistics. A summary of classification accuracy results for every state with a linking study is shown in Table 8.1.

The classification accuracy results show how well MAP Growth predicts whether students will meet proficiency levels on state summative assessments. Accuracy is reported by state, subject (reading and mathematics), and grade and reflects the percentage of students who were correctly classified as proficient or not proficient based on their MAP Growth scores.

Across states and grades, classification accuracy is consistently high, indicating a strong alignment between MAP Growth and state summative assessments. In most cases, accuracy rates fall in the mid-80% to mid-90% range, meaning that MAP Growth correctly predicts state test performance for a large majority of students.

Several clear patterns from these linking studies have emerged:

- Math generally shows slightly higher accuracy than Reading, particularly in the upper elementary and middle grades. Reading is lower because MAP Growth Reading only involves reading content whereas the state summative tests combine Reading and Language Usage.
- Accuracy tends to be strongest in grades 3–8, where MAP Growth and state tests measure closely aligned content.
- Higher grades often show equal or higher accuracy compared with lower grades, reflecting increased stability in student performance over time.
- Accuracy is consistent across states, even though state assessments differ in design and standards.

These results provide strong evidence that MAP Growth is a reliable tool for identifying students who are on track to meet state expectations, as well as those who may need additional support. Because classification accuracy is reported separately by grade and subject, educators can have confidence that the projections are appropriate for their specific context.

Overall, the classification accuracy findings support the use of MAP Growth for instructional planning, early intervention, and data-informed decision-making well before spring state testing.

Table 8.1. Classification Accuracy by Subject and Grade for States with a Linking Study

Subject	State	Grade								
		3	4	5	6	7	8	9	10	11
Math	AK	0.95	0.95	0.94	0.95	0.96	0.95	0.96	–	–
	AR	0.84	0.84	0.84	0.84	0.87	0.85	0.88	0.89	–
	AZ	0.88	0.87	0.87	0.89	0.89	0.89	–	–	–
	CA	0.87	0.89	0.89	0.89	0.91	0.91	–	–	0.89
	CO	0.86	0.88	0.89	0.91	0.90	0.89	–	–	–
	FL	0.89	0.88	0.88	0.84	0.80	0.78	–	–	–
	GA	0.88	0.88	0.88	0.89	0.89	0.84	–	–	–
	IA	0.87	0.89	0.88	0.87	0.87	0.85	0.87	0.86	–
	IL	0.85	0.86	0.88	0.89	0.88	0.88	–	–	–
	IN	0.88	0.87	0.88	0.87	0.89	0.88	–	–	–
	KS	0.87	0.88	0.89	0.89	0.89	0.90	–	–	–
	KY	0.85	0.85	0.86	0.85	0.83	0.83	–	0.81	–
	MA	0.83	0.85	0.87	0.87	0.90	0.88	–	–	–
	MD	0.90	0.91	0.89	0.93	0.92	0.95	–	–	–
	MI	0.86	0.87	0.89	0.90	0.90	0.87	–	–	–
	MN	0.90	0.90	0.89	0.90	0.90	0.90	–	–	–
	MO	0.86	0.86	0.88	0.88	0.88	0.87	–	–	–
	MS	0.85	0.85	0.85	0.87	0.87	0.86	–	–	–
	NC	0.87	0.87	0.88	0.86	0.88	0.82	–	–	–
	ND	0.84	0.84	0.87	0.86	0.88	0.85	–	–	–
	NE	0.94	0.96	0.96	0.95	0.96	0.95	–	–	–
	NJ	0.87	0.87	0.89	0.90	0.89	0.91	–	–	–
	NM	0.88	0.88	0.86	0.85	0.88	0.88	–	–	–
	NV	0.87	0.86	0.89	0.89	0.90	0.90	–	–	–
	NY	0.85	0.87	0.86	0.87	0.86	0.83	–	–	–
	OH	0.87	0.88	0.87	0.88	0.87	0.82	–	–	–
	OK	0.84	0.85	0.88	0.87	0.87	0.90	–	–	–
	OR	0.87	0.88	0.89	0.87	0.89	0.89	–	–	–
	PA	0.88	0.88	0.87	0.88	0.91	0.89	–	–	–
	SBAC	0.87	0.87	0.89	0.89	0.90	0.91	–	–	–
	SC	0.88	0.88	0.88	0.89	0.90	0.89	–	–	–
	SD	0.86	0.87	0.88	0.87	0.88	0.87	–	–	–
	TN	0.86	0.86	0.87	0.87	0.87	0.86	–	–	–
	TX	0.86	0.87	0.85	0.88	0.87	0.82	–	–	–
	VA	0.90	0.91	0.90	0.89	0.87	0.85	–	–	–
	WA	0.87	0.88	0.88	0.88	0.89	0.88	–	–	–

Subject	State	Grade								
		3	4	5	6	7	8	9	10	11
	WI	0.86	0.87	0.87	0.90	0.89	0.87	–	–	–
Reading	AK	0.93	0.92	0.93	0.94	0.93	0.93	0.94	–	–
	AR	0.84	0.84	0.82	0.81	0.81	0.80	0.83	0.82	–
	AZ	0.85	0.85	0.87	0.86	0.85	0.86	–	–	–
	CA	0.85	0.85	0.85	0.85	0.85	0.86	–	–	0.84
	CO	0.84	0.85	0.82	0.84	0.83	0.82	–	–	–
	FL	0.84	0.84	0.84	0.85	0.84	0.85	0.86	0.84	–
	GA	0.87	0.86	0.85	0.85	0.85	0.82	0.87	0.86	–
	IA	0.86	0.86	0.85	0.85	0.87	0.85	0.89	0.87	–
	IL	0.81	0.81	0.82	0.82	0.80	0.79	–	–	–
	IN	0.83	0.83	0.82	0.83	0.82	0.82	–	–	–
	KS	0.86	0.85	0.86	0.85	0.84	0.87	–	–	–
	KY	0.82	0.82	0.83	0.82	0.82	0.81	–	0.81	–
	MA	0.81	0.81	0.82	0.83	0.83	0.81	–	–	–
	MD	0.87	0.87	0.87	0.85	0.86	0.86	–	–	–
	MI	0.84	0.85	0.85	0.85	0.85	0.77	–	–	–
	MN	0.86	0.86	0.88	0.87	0.87	0.85	–	–	–
	MO	0.84	0.83	0.83	0.83	0.83	0.82	–	–	–
	MS	0.83	0.82	0.84	0.85	0.84	0.85	–	–	–
	NC	0.84	0.84	0.84	0.84	0.83	0.83	–	–	–
	ND	0.79	0.81	0.82	0.81	0.80	0.82	–	–	–
	NE	0.93	0.94	0.94	0.94	0.92	0.93	–	–	–
	NJ	0.82	0.83	0.82	0.82	0.80	0.8	–	–	–
	NM	0.87	0.85	0.86	0.84	0.82	0.82	–	–	–
	NV	0.84	0.85	0.85	0.85	0.85	0.84	–	–	–
	NY	0.83	0.83	0.82	0.83	0.81	0.82	–	–	–
	OH	0.82	0.84	0.84	0.82	0.83	0.81	–	–	–
	OK	0.82	0.83	0.83	0.83	0.83	0.83	–	–	–
	OR	0.85	0.83	0.85	0.83	0.85	0.85	–	–	–
	PA	0.86	0.86	0.85	0.84	0.83	0.83	–	–	–
	SBAC	0.86	0.85	0.85	0.84	0.84	0.85	–	–	–
	SC	0.86	0.86	0.87	0.86	0.86	0.85	–	–	–
	SD	0.83	0.83	0.84	0.86	0.84	0.83	–	–	–
	TN	0.83	0.83	0.82	0.85	0.83	0.82	–	–	–
	TX	0.80	0.82	0.83	0.82	0.84	0.83	–	–	–
	VA	0.83	0.86	0.85	0.84	0.83	0.84	–	–	–
	WA	0.87	0.85	0.86	0.83	0.84	0.84	–	–	–
	WI	0.84	0.86	0.86	0.85	0.84	0.84	–	–	–

Subject	State	Grade								
		3	4	5	6	7	8	9	10	11
Language Usage	KY	–	–	0.78	–	–	0.80	–	–	–
Physical Science	GA	–	–	–	–	–	0.81	–	–	–
Science	FL	–	–	0.84	–	–	0.86	–	–	–
	GA	–	–	0.86	–	–	0.85	–	–	–
	IN	–	0.8	–	0.84	–	–	–	–	–
	KY	–	0.81	–	–	0.87	–	–	–	–
	MI	–	–	0.84	–	–	0.84	–	–	–
	MO	–	–	0.85	–	–	0.84	–	–	–
	NJ	–	–	0.89	–	–	0.93	–	–	–
	NY	–	–	0.81	–	–	–	–	–	–
	OH	–	–	0.82	–	–	0.82	–	–	–
	TX	–	–	0.84	–	–	0.82	–	–	–

8.4. Relationships with Measures of the Same Construct

In addition to classification accuracy, the correlation between MAP Growth and state summative assessment scores is also calculated in a linking study. Correlations for all states with a linking study are shown in Table 8.2.

The correlations describe how closely MAP Growth scores and summative results tend to vary together. In general, higher correlations mean that students who perform well on MAP Growth also tend to perform well on the summative tests, while lower correlations indicate a weaker relationship between the two measures. Across the states included in the table, correlations are consistently strong, which supports using MAP Growth as a meaningful indicator of performance on state summative assessments—especially when interpreted alongside classroom evidence and other local measures.

Overall, the relationship is stronger in Math than in Reading. Averaged across states and the grades available in the dataset, Math correlations are about 0.86, compared with about 0.82 for Reading. Put simply, MAP Growth tends to track summative outcomes a little more closely in Math than it does in Reading in these results. This does not mean Reading is weak; rather, both subjects show strong relationships, with Math being modestly higher on average.

Looking across grade levels where coverage is most complete (grades 3–8), the pattern is fairly consistent. Correlations are strongest in the earlier grades—particularly grades 3 through 6—where the average relationship is high and stable for both subjects. In Math, the average correlation peaks around grade 6 then decreases somewhat by grade 8. Reading follows a similar but slightly flatter pattern, remaining relatively steady through the middle grades and then dipping modestly by grade 8. While the file includes some values for grades 9–11, those appear for only a small subset of states, so those later-grade results should be interpreted cautiously and not treated as representative of broad patterns in the same way as grades 3–8.

The table also includes correlations for Science, Physical Science, and Language Usage, but those entries are much more limited than Math and Reading. Because they represent far fewer

state-grade combinations, they are best viewed as supplemental information rather than a basis for broad conclusions about alignment in those subjects.

Taken together, these correlations reinforce a key message for educators: MAP Growth provides a strong signal related to summative performance, particularly in grades 3–6, and tends to be slightly stronger in Math than in Reading.

Table 8.2. Correlations Between MAP Growth and State Summative Assessments

Subject	State	Grade								
		3	4	5	6	7	8	9	10	11
Math	AK	0.96	0.96	0.95	0.97	0.96	0.96	0.96	–	–
	AR	0.85	0.82	0.80	0.81	0.85	0.85	0.84	0.86	–
	AZ	0.87	0.89	0.88	0.91	0.86	0.87	–	–	–
	CA	0.88	0.90	0.90	0.92	0.93	0.92	–	–	0.89
	CO	0.86	0.87	0.86	0.89	0.87	0.86	–	–	–
	FL	0.89	0.88	0.89	0.85	0.73	0.69	–	–	–
	GA	0.87	0.87	0.87	0.87	0.84	0.80	–	–	–
	IA	0.83	0.87	0.86	0.85	0.85	0.84	0.85	0.84	–
	IL	0.84	0.85	0.85	0.84	0.86	0.84	–	–	–
	IN	0.89	0.90	0.91	0.90	0.90	0.89	–	–	–
	KS	0.84	0.85	0.85	0.84	0.84	0.84	–	–	–
	KY	0.80	0.81	0.81	0.79	0.76	0.77	–	0.71	–
	MA	0.82	0.85	0.86	0.86	0.85	0.83	–	–	–
	MD	0.88	0.87	0.86	0.91	0.85	0.74	–	–	–
	MI	0.87	0.88	0.89	0.92	0.91	0.86	–	–	–
	MN	0.92	0.92	0.93	0.93	0.93	0.92	–	–	–
	MO	0.86	0.84	0.86	0.87	0.87	0.83	–	–	–
	MS	0.85	0.85	0.85	0.89	0.85	0.87	–	–	–
	NC	0.82	0.83	0.87	0.87	0.86	0.75	–	–	–
	ND	0.84	0.87	0.89	0.88	0.88	0.86	–	–	–
	NE	0.98	0.98	0.99	0.99	0.99	0.99	–	–	–
	NJ	0.86	0.88	0.88	0.88	0.88	0.76	–	–	–
	NM	0.81	0.82	0.79	0.76	0.76	0.75	–	–	–
	NV	0.89	0.90	0.89	0.91	0.90	0.88	–	–	–
	NY	0.82	0.84	0.83	0.86	0.85	0.79	–	–	–
	OH	0.85	0.86	0.84	0.85	0.84	0.77	–	–	–
	OK	0.83	0.83	0.85	0.85	0.82	0.83	–	–	–
	OR	0.89	0.91	0.90	0.90	0.90	0.88	–	–	–
	PA	0.85	0.87	0.86	0.84	0.86	0.86	–	–	–
	SBAC	0.88	0.89	0.90	0.91	0.92	0.91	–	–	–
	SC	0.85	0.86	0.87	0.86	0.86	0.86	–	–	–
	SD	0.89	0.90	0.90	0.91	0.92	0.90	–	–	–
	TN	0.84	0.85	0.86	0.86	0.85	0.83	–	–	–

Subject	State	Grade								
		3	4	5	6	7	8	9	10	11
	TX	0.79	0.81	0.80	0.81	0.80	0.76	–	–	–
	VA	0.84	0.83	0.84	0.85	0.78	0.80	–	–	–
	WA	0.88	0.90	0.91	0.91	0.92	0.91	–	–	–
	WI	0.86	0.88	0.86	0.89	0.86	0.84	–	–	–
Reading	AK	0.94	0.93	0.94	0.94	0.94	0.94	0.94	–	–
	AR	0.80	0.81	0.80	0.79	0.79	0.79	0.76	0.78	–
	AZ	0.83	0.85	0.85	0.84	0.85	0.83	–	–	–
	CA	0.85	0.85	0.86	0.85	0.86	0.85	–	–	0.82
	CO	0.84	0.82	0.81	0.82	0.81	0.78	–	–	–
	FL	0.78	0.82	0.83	0.83	0.80	0.80	0.79	0.78	–
	GA	0.87	0.85	0.85	0.85	0.83	0.78	0.87	0.85	–
	IA	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	–
	IL	0.78	0.77	0.77	0.77	0.76	0.73	–	–	–
	IN	0.82	0.82	0.82	0.81	0.81	0.81	–	–	–
	KS	0.85	0.84	0.83	0.82	0.80	0.81	–	–	–
	KY	0.73	0.75	0.77	0.78	0.77	0.76	–	0.73	–
	MA	0.78	0.79	0.78	0.77	0.78	0.77	–	–	–
	MD	0.86	0.86	0.86	0.85	0.85	0.85	–	–	–
	MI	0.83	0.84	0.83	0.83	0.83	0.77	–	–	–
	MN	0.89	0.89	0.89	0.88	0.87	0.86	–	–	–
	MO	0.84	0.82	0.82	0.82	0.81	0.81	–	–	–
	MS	0.81	0.79	0.82	0.81	0.82	0.81	–	–	–
	NC	0.83	0.82	0.81	0.78	0.80	0.79	–	–	–
	ND	0.77	0.79	0.80	0.81	0.77	0.82	–	–	–
	NE	0.97	0.97	0.97	0.97	0.97	0.97	–	–	–
	NJ	0.81	0.81	0.80	0.80	0.78	0.78	–	–	–
	NM	0.86	0.84	0.81	0.82	0.79	0.76	–	–	–
	NV	0.83	0.85	0.83	0.83	0.83	0.83	–	–	–
	NY	0.81	0.81	0.80	0.81	0.80	0.80	–	–	–
	OH	0.77	0.81	0.79	0.80	0.78	0.79	–	–	–
	OK	0.82	0.83	0.82	0.81	0.81	0.81	–	–	–
	OR	0.87	0.86	0.86	0.85	0.84	0.83	–	–	–
	PA	0.84	0.84	0.84	0.81	0.81	0.79	–	–	–
	SBAC	0.85	0.85	0.85	0.85	0.85	0.84	–	–	–
	SC	0.87	0.85	0.85	0.85	0.84	0.84	–	–	–
	SD	0.86	0.84	0.85	0.86	0.84	0.80	–	–	–
	TN	0.78	0.81	0.79	0.82	0.78	0.80	–	–	–
	TX	0.73	0.73	0.77	0.76	0.79	0.75	–	–	–
	VA	0.78	0.79	0.79	0.77	0.75	0.76	–	–	–

Subject	State	Grade								
		3	4	5	6	7	8	9	10	11
	WA	0.85	0.86	0.85	0.84	0.85	0.84	–	–	–
	WI	0.84	0.83	0.84	0.83	0.83	0.82	–	–	–
Language Usage	KY	–	–	0.69	–	–	0.74	–	–	–
Physical Science	GA	–	–	–	–	–	0.73	–	–	–
Science	FL	–	–	0.84	–	–	0.82	–	–	–
	GA	–	–	0.84	–	–	0.76	–	–	–
	IN	–	0.79	–	0.82	–	–	–	–	–
	KY	–	0.62	–	–	0.72	–	–	–	–
	MI	–	–	0.83	–	–	0.83	–	–	–
	MO	–	–	0.83	–	–	0.83	–	–	–
	NJ	–	–	0.84	–	–	0.80	–	–	–
	NY	–	–	0.71	–	–	–	–	–	–
	OH	–	–	0.71	–	–	0.73	–	–	–
	TX	–	–	0.78	–	–	0.77	–	–	–

8.5. Universal Screening

Universal screening is paramount in identifying students at risk for academic difficulty in a response to intervention (RTI) model, the core of which is to provide students with multi-tiered support based on the level of academic risk that students encounter. Typically, a multitiered support system consists of three levels (Tier 1, Tier 2, and Tier 3) escalating from no intervention needed to the most intense level of intervention. It is estimated that 5–10% of the student population needs the most intensive intervention.

One primary component in RTI is assessment. A universal screening assessment in a particular content domain is typically administered multiple times a year. If a student scores below an established benchmark for a given time point, they are considered at risk for learning difficulties in that content domain and in need of intervention. For an assessment to be an effective universal screener (aside from technical adequacy), it is imperative to establish benchmarks through a scientifically designed and evidenced-based process. The benchmarks also need to be explicit as to what level of academic risk they are established to identify (e.g., at some risk or at substantial risk).

He and Meyer (2021) established the universal screening cut scores for the English MAP Growth assessments pursuant to the NCII rating rubrics (NCII, 2020). The primary sample consisted of students in Arkansas, Colorado, Florida, Missouri, and New York, and a secondary sample used for cross-validation consisted of students in Indiana. The primary sample took state-level summative tests and MAP Growth in Spring 2018, whereas the secondary sample took the state summative test and MAP Growth in Spring 2019. While the original Indiana state assessment scale scores were used as the criterion measure in the classification accuracy analyses for the secondary sample, state assessment scores from the primary sample were put on the same scale (i.e., the RIT scale) by subject and grade using the equipercentile method to create a common criterion measure and allow state-level test scores to be comparable across

states. As a result, each student in the primary sample obtained a MAP Growth linked state score in Math and/or Reading.

Thresholds for identifying students in need of intervention were based the classification accuracy analyses. Thresholds were chosen to maximize classification accuracy. Students who score below those thresholds are likely at risk for severe learning difficulty and in need of intensive intervention. The thresholds resulted in sensitivity, specificity, and the lower bound of the area under curve (AUC) being at least 0.8, the highest level of the evaluation criteria described in the NCII rating rubrics (NCII, 2020). The cross-validation study results were consistent with those from the primary sample, providing evidence that the recommended universal screening cut scores are valid.

NWEA updated the MAP Growth universal screening thresholds in Math and Reading in 2025 to address score adjustments introduced with the Enhanced Item Selection Algorithm (EISA) and to align with the newly released 2025 MAP Growth national norms (this was an update to He & Meyer, 2021, to be formally published in 2026). The new study used the same sample and procedures as the 2021 study, with slight modifications for Math, including converting traditional MAP Growth scores to EISA-based scores and applying the equipercentile method for deriving the candidate criterion measure cut scores needed by the classification accuracy analysis. Details on the thresholds from this study are provided in Table 8.3.

Following the same NCII rating rubrics (NCII, 2020), the study finds that, for both subjects, MAP Growth thresholds at the 35th percentile on the 2025 norms provide the best balance between sensitivity and specificity for identifying students truly in need of intensive intervention. Students whose scores fall below these thresholds are likely at risk of having learning difficulties in the respective subject areas.

Table 8.3. Universal Screening Thresholds

Grade	Term	Math		Reading	
		Threshold	Percentile	Threshold	Percentile
K	Fall	136	35	134	35
	Winter	146	35	142	35
	Spring	153	35	147	35
1	Fall	154	35	150	35
	Winter	163	35	157	35
	Spring	169	35	162	35
2	Fall	167	35	163	35
	Winter	175	35	170	35
	Spring	181	35	175	35
3	Fall	178	35	178	35
	Winter	186	35	183	35
	Spring	193	35	187	35
4	Fall	191	35	189	35
	Winter	198	35	193	35
	Spring	203	35	195	35
5	Fall	200	35	197	35
	Winter	205	35	200	35

Grade	Term	Math		Reading	
		Threshold	Percentile	Threshold	Percentile
	Spring	209	35	202	35
6	Fall	204	35	202	35
	Winter	209	35	204	35
	Spring	213	35	206	35
7	Fall	211	35	206	35
	Winter	214	35	207	35
	Spring	216	35	208	35
8	Fall	215	35	209	35
	Winter	219	35	210	35
	Spring	221	35	211	35

8.6. The Learning Continuum

8.6.1. 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 that an item measures. All items assessing the same skill or concept are aligned to the same statement. Because a MAP Growth item pool contains thousands of items across grades and domains, 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 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.

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

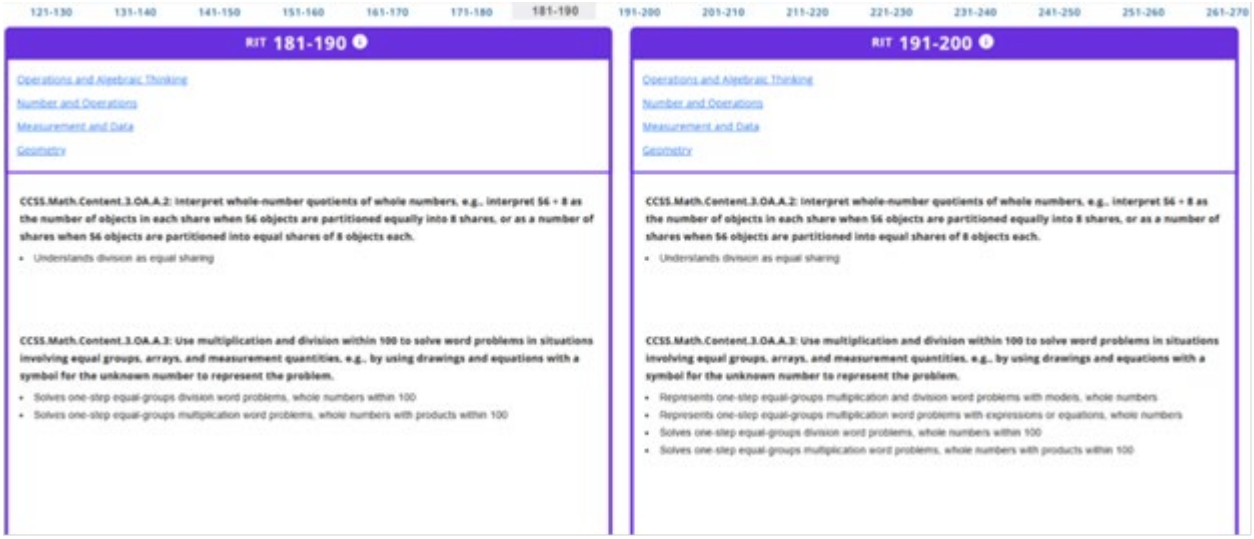
The Learning Continuum and its Learning Statements represent the content domain underlying the MAP Growth RIT scale by describing the skills and concepts reflected in the item pool at different points along the scale. Learning Statements provide instructionally relevant descriptors of what test items measure, and grouping these statements by RIT range illustrates the types of content typically associated with each 10-point band. Importantly, this linkage helps clarify what a student's RIT score represents in terms of the underlying content domain without prescribing instructional placement, defining mastery, or focusing only on grade-level expectations. Instead, the Learning Continuum provides students, educators, assessment designers, and curriculum specialists with insight into how items map onto the scale and how student performance aligns with specific content features within the larger domain structure.

8.6.3. Instructional Uses of the Learning Continuum

Educators can use the Learning Continuum to deepen their understanding of the types of content represented on 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 supports educators in identifying skills and concepts students may have been assessed on and in coordinating MAP Growth data with standards and curriculum materials during term or unit planning. It 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. While MAP Growth offers value for planning at

broader instructional intervals, high-quality formative assessment remains essential for tailoring daily instruction. Used together, MAP Growth data, the Learning Continuum, and ongoing formative practices provide a complementary set of tools for supporting instructional decision-making, understanding instructional impact, and promoting student learning.

Figure 8.1. Learning Continuum Example



9. The Test Is Fair for All Examinees

9.1. Test Taker Demographics

The MAP Growth 2024–2025 assessments were administered to a large and diverse population of students across subjects and testing terms. Across Math and Reading, examinee counts exceeded 8.5 million students per term, while Language Usage and Science assessments included between approximately 1.2 and 1.8 million students per term. These large sample sizes provide strong support for the stability and generalizability of reported results across demographic subgroups. Demographic information is reported for student gender and race/ethnicity because it is the only demographic information that school districts are required to submit. Other demographic information such as English language learner and student disability status is voluntary and not consistently reported by school districts.

As shown in Table 9.1, gender representation among MAP Growth examinees was highly balanced across all subjects and testing terms. Male students comprised approximately 50.7% to 51.1% of examinees, while female students comprised approximately 48.9% to 49.3%. This near parity was consistent across Math, Reading, Language Usage, and Science and showed minimal variation across fall, winter, and spring administrations.

Table 9.1. Demographics by Gender

Subject	Term	Total	Percentage Male	Percentage Female
Math	Fall	9,228,120	50.84	49.16
Math	Winter	8,502,422	50.86	49.14
Math	Spring	8,557,420	50.79	49.21
Reading	Fall	9,737,063	51.02	48.98
Reading	Winter	8,840,778	51.06	48.94
Reading	Spring	8,833,663	50.96	49.04
Language Usage	Fall	1,572,412	50.75	49.25
Language Usage	Winter	1,261,942	50.76	49.24
Language Usage	Spring	1,404,652	50.67	49.33
Science	Fall	1,835,803	50.80	49.20
Science	Winter	1,555,054	50.90	49.10
Science	Spring	1,723,645	50.78	49.22

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 was generally consistent across subjects and testing seasons. White students constituted the largest proportion of examinees, representing approximately 41% to 46% of the population in Math and Reading and a slightly lower proportion (about 36% to 37%) in Science. Hispanic students comprised the second-largest group, accounting for roughly 25% of Math and Reading examinees and approximately 30% of Science examinees. Black students represented approximately 15% of examinees across Math, Reading, and Science, while Asian students accounted for about 5% across most subjects. Students identifying as American Indian or Alaska Native and Native Hawaiian or Pacific Islander each represented smaller proportions of the population (generally 1%–2% and less than 0.5%, respectively). Students identifying as Other race or Multiple races together accounted for approximately 9%–12% of examinees, with somewhat higher representation in the Language Usage assessment. Overall, the racial and ethnic distributions were stable across fall, winter, and spring administrations within each subject.

Appendix B provides detailed demographic distributions of students who participated in the 2024–2025 MAP Growth administration by subject, grade level, and testing term. While the tables are presented in full in the appendix due to their length and granularity, a summary of key patterns is provided here to support the interpretation of results.

Across all subjects, grades, and testing terms, gender representation among MAP Growth examinees was highly balanced. In nearly every grade from kindergarten through grade 12, the proportion of male students ranged from approximately 50% to 52%, with a corresponding proportion of female students ranging from approximately 48% to 50%.

In Math and Reading, slight increases in the percentage of male students were observed at higher grade levels, particularly in grades 9–12, where males represented approximately 51% to 53% of examinees. Similar patterns were observed in Language Usage and Science, though these subjects included smaller overall sample sizes at the upper grades. Across subjects, differences in gender representation across fall, winter, and spring administrations within the same grade were minimal, typically less than one percentage point. Overall, the results indicate a stable and nearly equal gender distribution across grades and terms.

According to the detailed demographic results in Appendix B, the racial and ethnic composition of MAP Growth examinees varied systematically by grade level and subject while remaining generally stable across testing terms within grades. In Math and Reading, White students represented the largest racial/ethnic group at most grade levels, accounting for approximately 40% to 43% of examinees in elementary and middle grades. At higher grade levels (grades 9–12), the proportion of White students decreased, while the proportions of Hispanic students increased, particularly in Math and Reading, where Hispanic representation exceeded 30% in several high school grades.

Hispanic students consistently represented the second-largest group across grades and subjects. In elementary grades, Hispanic students accounted for approximately 23% to 26% of Math and Reading examinees, with higher proportions observed in secondary grades. Black students comprised approximately 14% to 16% of examinees in Math and Reading across most grades, with modest increases in some upper-grade Reading administrations. Asian students generally accounted for approximately 4% to 6% of examinees, with slightly higher representation in middle grades relative to early elementary and high school grades.

Students identifying as American Indian or Alaska Native and Native Hawaiian or Pacific Islander represented smaller proportions across all grades and subjects, generally ranging from about 1% to 3% and less than 1%, respectively. Students identifying as Other race or Multiple races together accounted for a small proportion of examinees, typically between 8% and 13%, with somewhat higher representation in Language Usage and Science assessments.

In Language Usage, the racial and ethnic composition differed somewhat from Math and Reading, with a higher proportion of White students (approximately 45%–49%) and lower proportions of Hispanic students in grades 2–8. At the high school level, however, Hispanic representation in Language Usage increased substantially, exceeding 40% in grade 12. Science assessments showed greater variability across grades, reflecting smaller sample sizes in early and upper secondary grades; nevertheless, Hispanic students represented a substantial proportion of examinees in Science, particularly in Grades 2–5 and grade 12.

Across all subjects and grades, racial and ethnic distributions were highly consistent across fall, winter, and spring testing terms, indicating stability in the demographic composition of

examinees throughout the academic year. Table 9.2 presents a condensed version of the results presented in full in Appendix B.

Table 9.2. Demographics by Race/Ethnicity

Subject	Term	Total	Percentage							
			Asian	Black	NH/PI	AI/AN	Hispanic	White	Other	Multiple
Math	Fall	9,228,127	5.22	14.91	0.28	1.42	25.09	41.68	6.37	5.03
Math	Winter	8,502,429	5.26	15.38	0.27	1.42	25.71	41.44	5.50	5.03
Math	Spring	8,557,427	5.35	15.03	0.26	1.30	25.19	41.25	6.29	5.33
Reading	Fall	9,737,070	5.30	15.48	0.29	1.36	25.06	41.35	6.22	4.94
Reading	Winter	8,840,786	5.34	15.93	0.29	1.34	25.45	41.21	5.49	4.96
Reading	Spring	8,833,670	5.40	15.44	0.27	1.25	25.14	40.98	6.26	5.27
Language	Fall	1,572,413	3.70	12.34	0.23	2.41	19.31	45.60	11.56	4.85
Language	Winter	1,261,942	4.14	13.71	0.24	2.68	19.06	45.12	10.32	4.74
Language	Spring	1,404,653	4.25	13.08	0.22	2.10	17.89	45.30	12.09	5.06
Science	Fall	1,835,808	5.08	15.31	0.26	2.15	30.39	37.07	5.04	4.70
Science	Winter	1,555,060	5.20	15.86	0.26	2.14	30.69	36.67	4.55	4.64
Science	Spring	1,723,650	5.33	15.52	0.25	1.95	30.46	36.17	5.24	5.09

Note. Examinees missing demographic information are omitted from this analysis.
Language = Language Usage; 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.

UDL Principle	Description
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.

Feature	Description
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.
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.
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.

Accommodation	Description
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).
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).

Third-Party Software	Description
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.

$$\mathrm{OR_k} = \frac{A_k D_k}{B_k C_k} \tag{12}$$

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

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

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 (ΔMH), 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.

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

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.

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

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 ΔMH 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 assessments in Math, Reading, and Science across fall, winter, and spring administrations in 2024–2025. 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 all subjects, terms, and group comparisons, the large majority of items were classified as A-level (negligible DIF). In Math, approximately 85–94% of items fell into the A category (A+ and A− combined), depending on the focal group and term. Reading showed even higher proportions of A-level items, approximately 92–97% across gender and racial/ethnic comparisons. Science results were similar, with roughly 94–97% of items classified as A-level across administrations and groups. These results indicate that most items function comparably across groups and do not exhibit meaningful DIF.

B-level DIF (moderate DIF) accounted for a relatively small proportion of items. In Math, B-level DIF generally ranged from about 5% to 11%, with slightly higher percentages observed for Asian–White and Black–White comparisons than for gender comparisons. In Reading and Science, B-level DIF was typically lower, often between 2% and 6% of items. B-level DIF appeared in both directions (favoring focal or reference groups), with no consistent pattern across terms or subjects suggesting systematic bias against a particular group.

C-level DIF (large DIF) was rare across all analyses. In Math, the percentage of items classified as C-level DIF generally remained below 4% for all focal groups and terms, with Black–White comparisons showing the highest—but still small—proportions. In Reading, C-level DIF was consistently below 2% across all comparisons, and in Science it was typically below 2% as well. These low rates of C-level DIF indicate that very few items exhibit substantial differential functioning that would warrant serious concern or removal.

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

Table 9.9. DIF Classification Summary

Subject	Term	Focal Group	Reference Group	N Items	Percentage in DIF Class					
					A+	A-	B+	B-	C+	C-
Math	Fall	Female	Male	17,718	50.60	40.66	1.66	5.02	0.13	1.94
	Fall	Asian	White	15,069	50.34	39.42	6.30	2.44	0.90	0.60
	Fall	Black	White	14,964	41.13	44.30	4.90	5.75	1.71	2.21
	Fall	Hispanic	White	16,892	44.39	48.35	2.55	3.39	0.26	1.07
	Winter	Female	Male	18,015	51.67	40.40	1.49	4.61	0.13	1.69
	Winter	Asian	White	15,146	51.00	39.74	4.95	2.81	0.87	0.63
	Winter	Black	White	15,167	42.29	44.85	4.21	4.92	2.06	1.67
	Winter	Hispanic	White	17,305	46.28	46.99	2.08	3.13	0.47	1.06
	Spring	Female	Male	18,540	52.89	39.82	1.55	4.17	0.09	1.49
	Spring	Asian	White	15,406	50.58	40.56	4.76	2.73	0.62	0.76
	Spring	Black	White	15,632	43.53	43.97	4.20	4.28	2.46	1.56
	Spring	Hispanic	White	17,837	48.21	45.42	2.01	2.72	0.64	0.98
Reading	Fall	Female	Male	11,070	56.03	40.89	1.11	1.40	0.12	0.45
	Fall	Asian	White	10,738	51.61	39.94	3.10	2.93	1.57	0.84
	Fall	Black	White	10,876	41.56	54.22	0.75	2.62	0.40	0.44
	Fall	Hispanic	White	10,921	42.73	53.80	0.38	2.44	0.05	0.60
	Winter	Female	Male	11,134	57.82	39.17	1.18	1.28	0.06	0.49
	Winter	Asian	White	10,637	50.64	41.62	2.82	3.07	1.02	0.82
	Winter	Black	White	10,890	39.34	56.00	0.80	3.09	0.35	0.43
	Winter	Hispanic	White	10,954	41.46	54.44	0.46	2.89	0.03	0.72
	Spring	Female	Male	11,151	58.43	38.70	1.01	1.36	0.08	0.41
	Spring	Asian	White	10,410	50.45	42.46	2.58	2.65	1.05	0.81
	Spring	Black	White	10,787	39.60	56.22	0.68	2.82	0.27	0.42
	Spring	Hispanic	White	10,964	41.69	54.45	0.40	2.73	0.05	0.67
Science	Fall	Female	Male	5,285	50.75	43.56	1.51	3.14	0.17	0.87
	Fall	Asian	White	4,365	53.01	41.03	2.77	2.13	0.53	0.53
	Fall	Black	White	4,655	48.18	46.87	1.14	2.56	0.60	0.64
	Fall	Hispanic	White	4,774	47.05	50.27	0.61	1.61	0.29	0.17
	Winter	Female	Male	5,251	51.84	43.25	1.28	2.59	0.25	0.80

Subject	Term	Focal Group	Reference Group	N Items	Percentage in DIF Class					
					A+	A-	B+	B-	C+	C-
	Winter	Asian	White	4,332	50.30	43.33	2.91	2.49	0.46	0.51
	Winter	Black	White	4,640	48.06	46.01	1.47	2.22	1.31	0.93
	Winter	Hispanic	White	4,790	46.16	50.46	0.94	1.48	0.56	0.40
	Spring	Female	Male	5,292	52.85	41.80	1.51	2.82	0.23	0.79
	Spring	Asian	White	4,389	48.64	44.98	2.55	2.83	0.46	0.55
	Spring	Black	White	4,722	47.78	45.76	1.29	2.48	1.99	0.70
	Spring	Hispanic	White	4,719	46.75	48.87	1.04	1.99	0.91	0.45

10. Score Reports Facilitate the Interpretation and Use of Scores

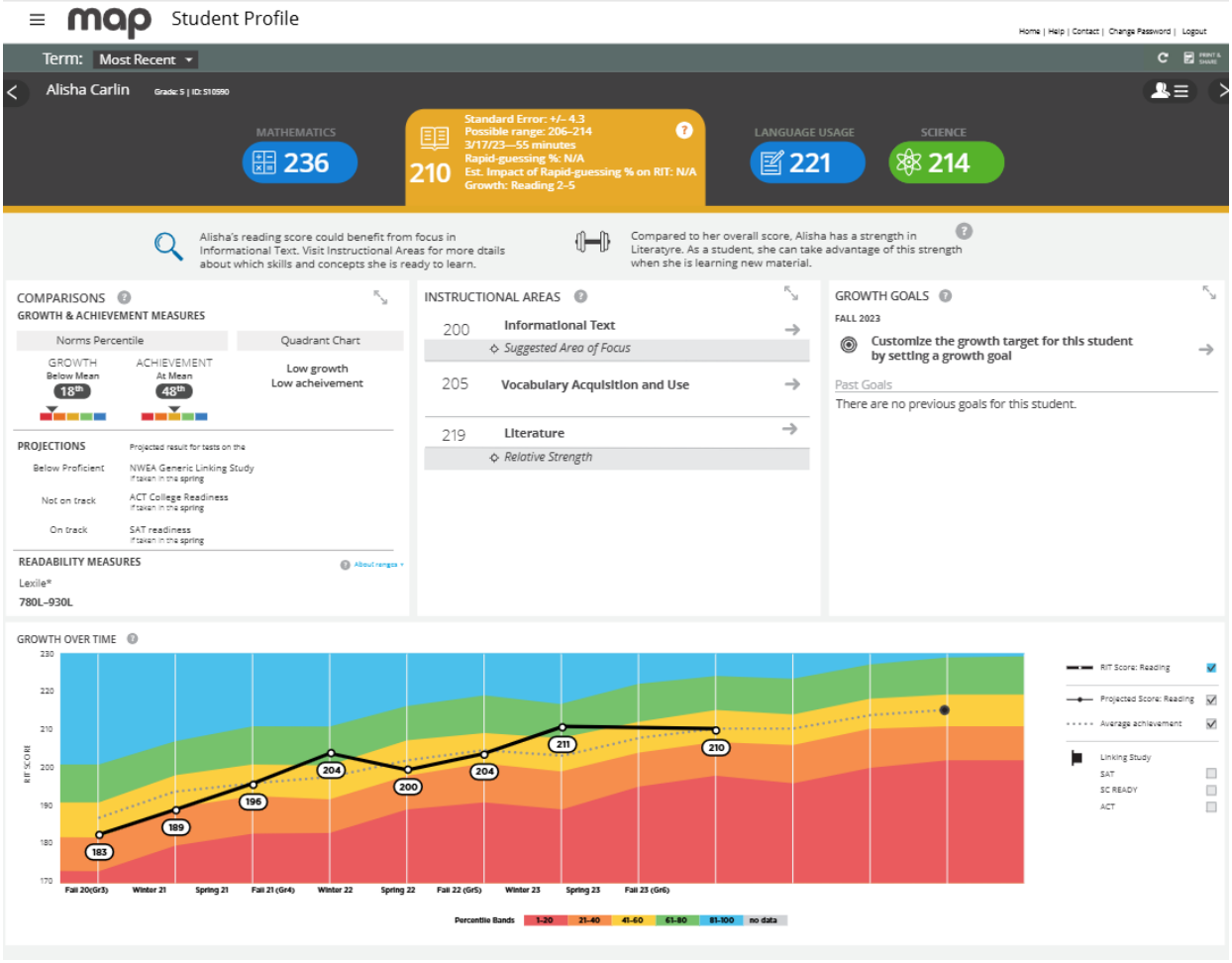
MAP Growth reports facilitate the interpretation of scores by clearly presenting information to a variety of users. MAP Growth reports are designed to support inferences at the student level and higher levels of aggregation, such as the classroom and district levels. Each report presents key information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

10.1. Student Profile Report

The Student Profile report, NWEA's most comprehensive student report, uses assessment data to create an individualized learner profile, allowing educators, caregivers, and students to track student performance throughout the school year and across years. This report shows a wealth of data—including growth medians and distribution data, current and past overall RIT scores, scores for instructional areas, 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 share how a student is performing, develop a personalized instructional plan, and collaboratively set goals. In addition, national norms data enables educators to better contextualize growth between two testing periods and understand how students' growth compares with similar students across the nation. This report also provides proficiency projections for state summative assessments and college readiness exams. An example Student Profile report is presented in Figure 10.1.

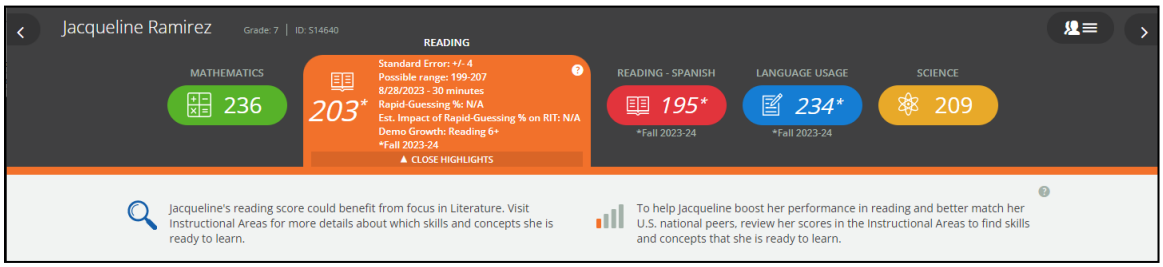
Figure 10.1. Example of Student Profile Report



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

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

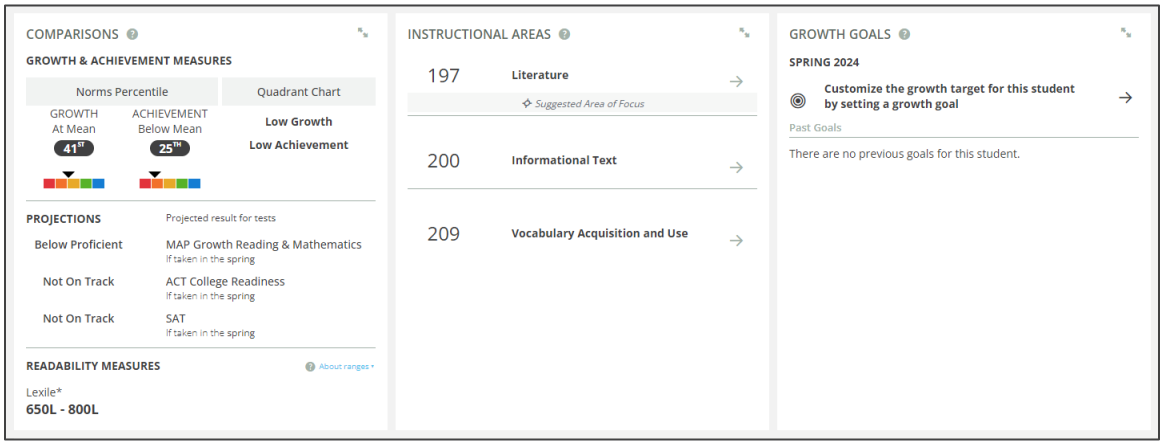
Figure 10.2. Example of Subject Scores and Highlights Sections of Student Profile Report



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

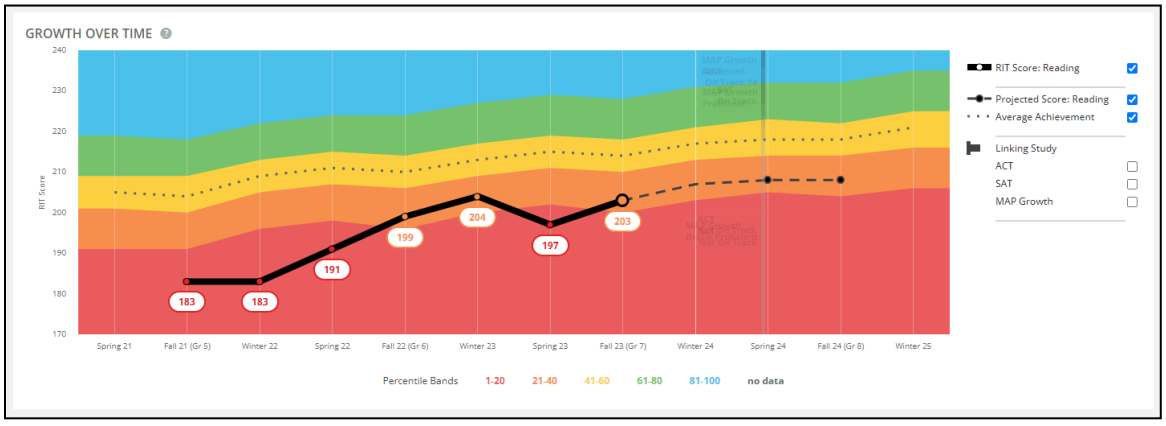
- The Comparisons section enables educators to place MAP Growth scores into a meaningful context. Information includes growth and achievement percentiles, projected proficiency for state summative assessments and college readiness exams, and Lexile or Quantile measures.
- The Instructional Areas section presents scores for each instructional area. Lower scores appear near the top so educators can identify where to focus efforts, and higher scores appear near the bottom so educators can celebrate students' strengths.
- The Growth Goals section enables educators and students to set a growth or performance target for each 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, as shown in Figure 10.4, displays student growth over time based on current and previous assessment results. It also provides predictive information based on typical growth and enables educators to enter specific goal information for a student.

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



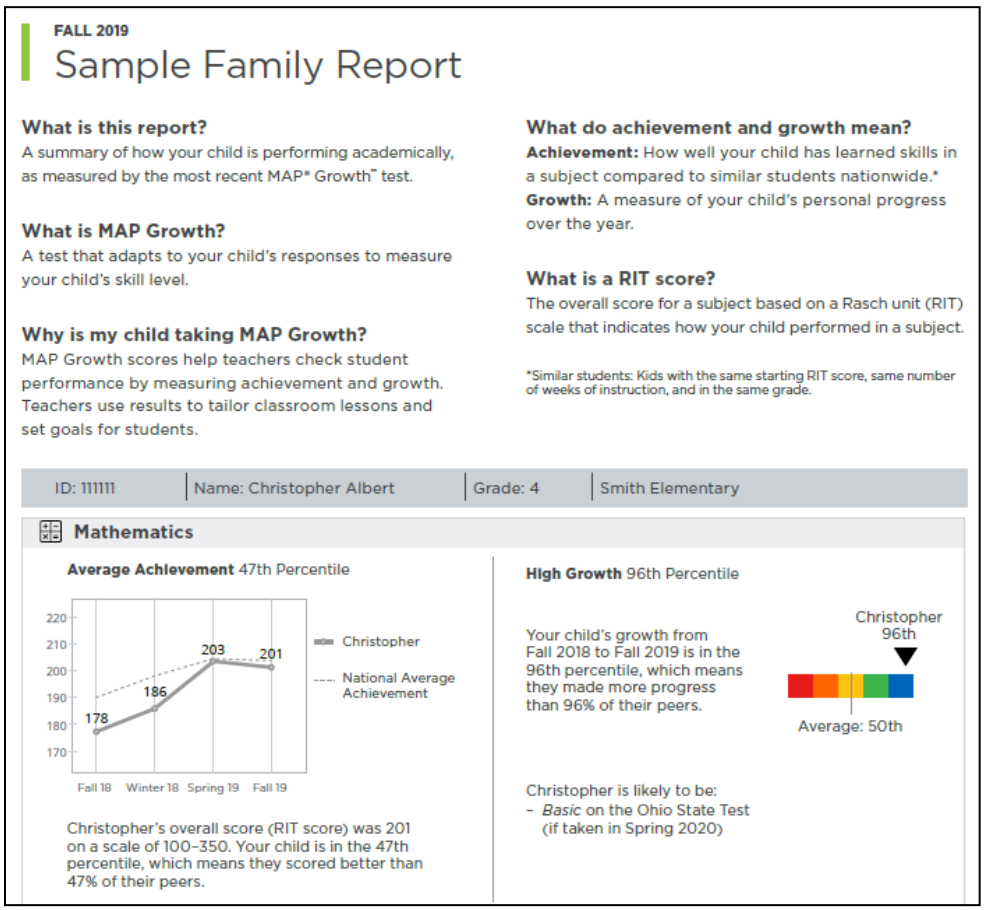
10.2. Family Report

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 easily share a 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 where a student might need extra support and how well they are doing in a subject compared with similar students nationwide. The report also includes questions educators, families and caregivers, and students can use to support meaningful, face-to-face conversations about assessments during conferences.

Educators and staff can batch-generate PDFs of reports by school, class, or individual student.

Figure 10.5. Example of Family Report



10.3. Class Reports

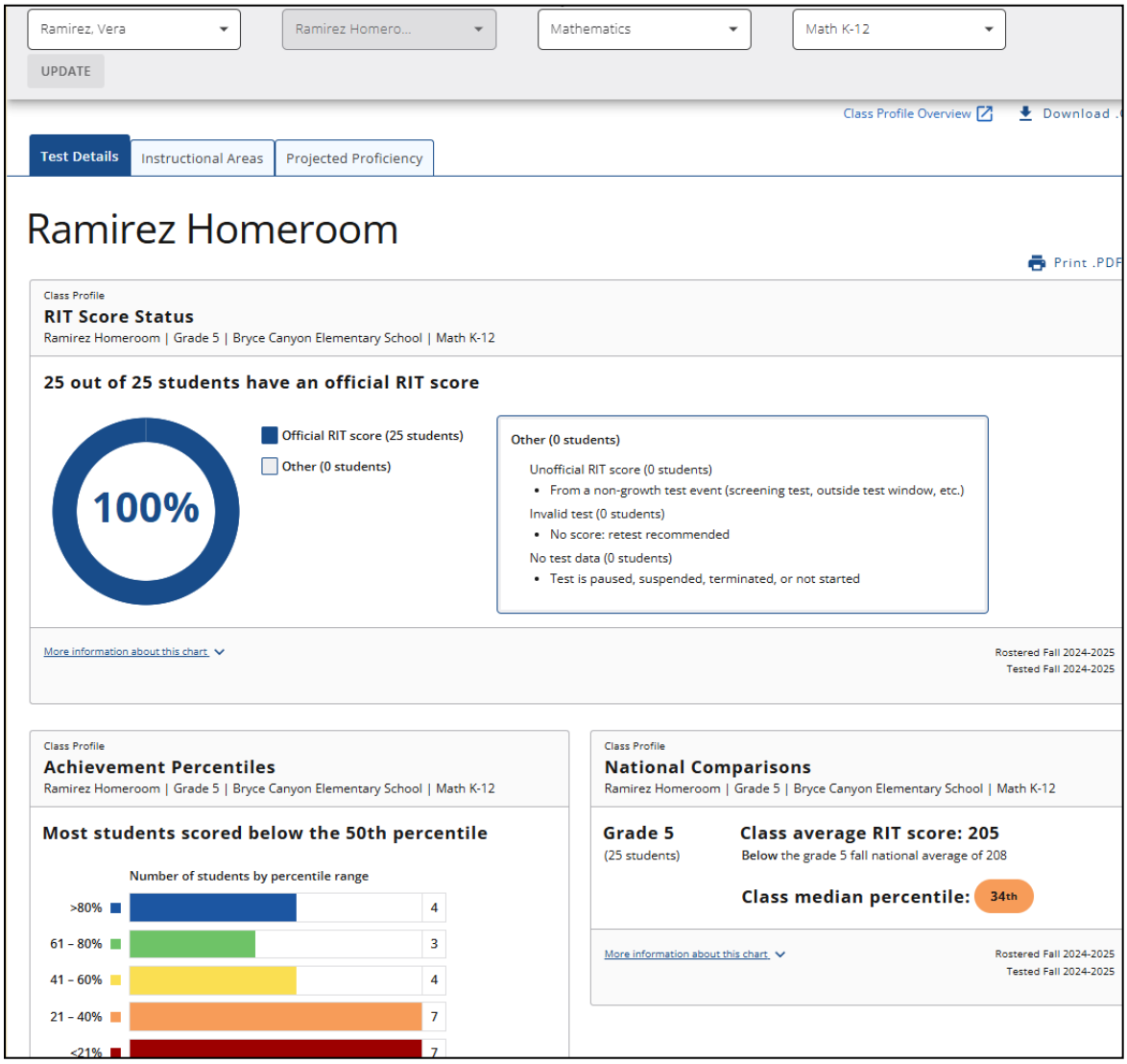
Reports at the class level provide an overview of performance and detailed information about each student in a class. Educators can use these reports to differentiate learning for individual students or groups of students, to inform classroom practice, and to identify instructional areas of strength and opportunity for the whole class.

10.3.1. Class Profile Report

The Class Profile report is an intuitive, interactive, and actionable report that provides class- and student-level insights to help educators plan instruction. Built with interactive features such as dynamic sorting and rapid access to the Student Profile report, the Class Profile report consolidates data from MAP Growth tests into easy-to-read views and offers decision-making tools to streamline educators' work.

As shown in the example presented in Figure 10.6, this report enables educators to identify valuable insights quickly and efficiently and informs decisions on "what's next" for their classes and/or individual students.

Figure 10.6. Example of Class Profile Report



Data visualizations provide a high-level overview of how the class performed across the instructional areas on the selected MAP Growth test and help educators see how students' instructional area scores are distributed across the five quintiles and 10-point RIT bands. This information allows teachers to gain a deeper understanding of areas of strength and opportunity for the class and provides a starting point as they make decisions about which areas of each

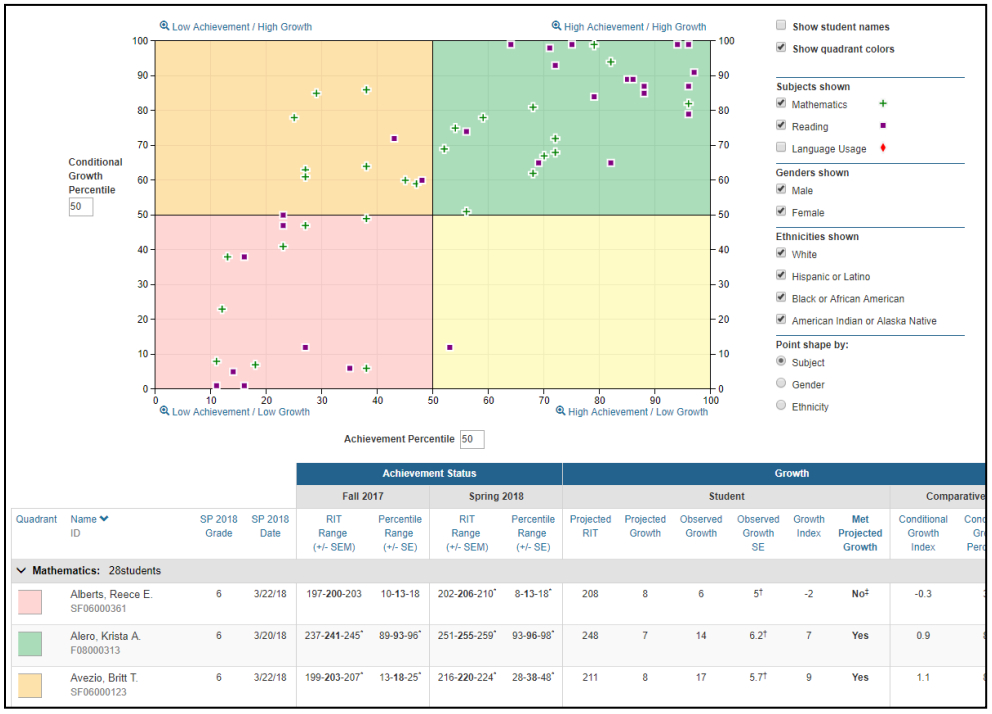
subject to emphasize during daily instruction. A 4-minute video about how educators can use the Class Profile report is provided at https://vimeo.com/537040943.

10.3.2. Achievement Status and Growth Report

The Achievement Status and Growth Report is particularly useful for measuring program effectiveness and student learning and for grouping students based on percentile information. This customizable report provides both a static and an interactive summary of data. The static report shows growth projections for each student (based on national norms) and compares actual student growth to projected growth.

With the interactive visualization, educators can quickly review how each student is growing and achieving so they can target instruction more effectively. The default setting characterizes achievement and growth relative to the 50th percentile, as presented in Figure 10.7.

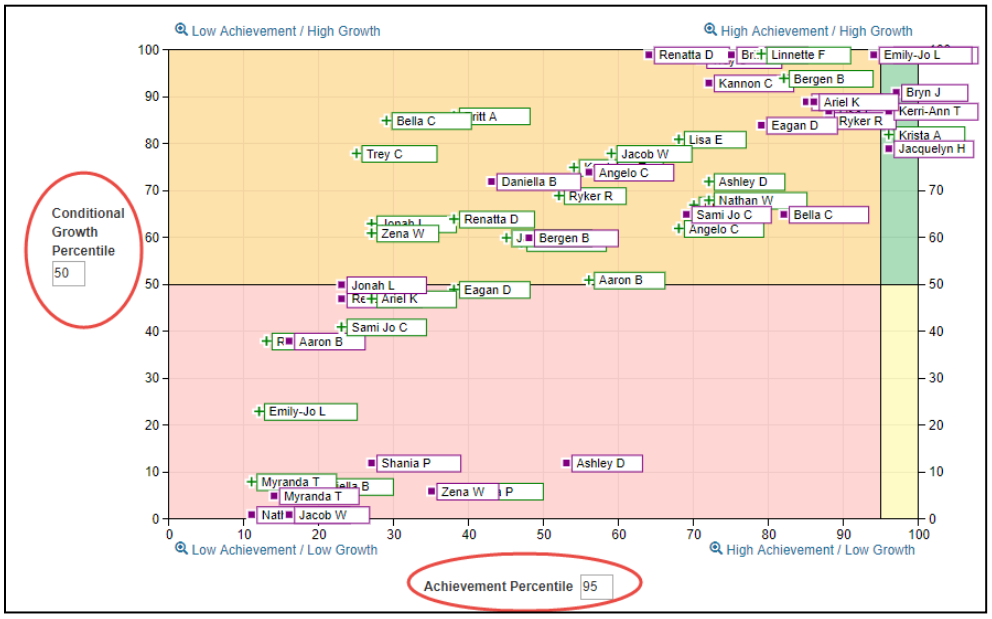
Figure 10.7. Example of Achievement Status and Growth Report



Quadrant	Name / ID	SP 2018 Grade	SP 2018 Date	Achievement Status: Fall 2017 RIT Range (+/- SEM)	Fall 2017 Percentile Range (+/- SE)	Spring 2018 RIT Range (+/- SEM)	Spring 2018 Percentile Range (+/- SE)	Growth: Student Projected RIT	Projected Growth	Observed Growth	Observed Growth SE	Growth Index	Met Projected Growth	Comparative: Conditional Growth Index	Conditional Growth Percentile
Mathematics: 28students															
	Alberts, Reece E. SF06000361	6	3/22/18	197-**200**-203	10-**13**-18	202-**206**-210*	8-**13**-18*	208	8	6	5†	-2	No‡	-0.3	
	Alero, Krista A. F08000313	6	3/20/18	237-**241**-245*	89-**93**-96*	251-**255**-259*	93-**96**-98*	248	7	14	6.2†	7	Yes	0.9	
	Avezio, Britt T. SF06000123	6	3/22/18	199-**203**-207*	13-**18**-25*	216-**220**-224*	28-**38**-48*	211	8	17	5.7†	9	Yes	1.1	

Using this report, educators can adjust the benchmarks against which achievement and growth are compared, as demonstrated in Figure 10.8, to group students for more effective instruction (e.g., intervention or extension) or for other purposes, such as identification 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 assess trends, identify areas of strength and opportunity, and review the percentage of students meeting targets, NWEA provides school- and district-level aggregated data reports.

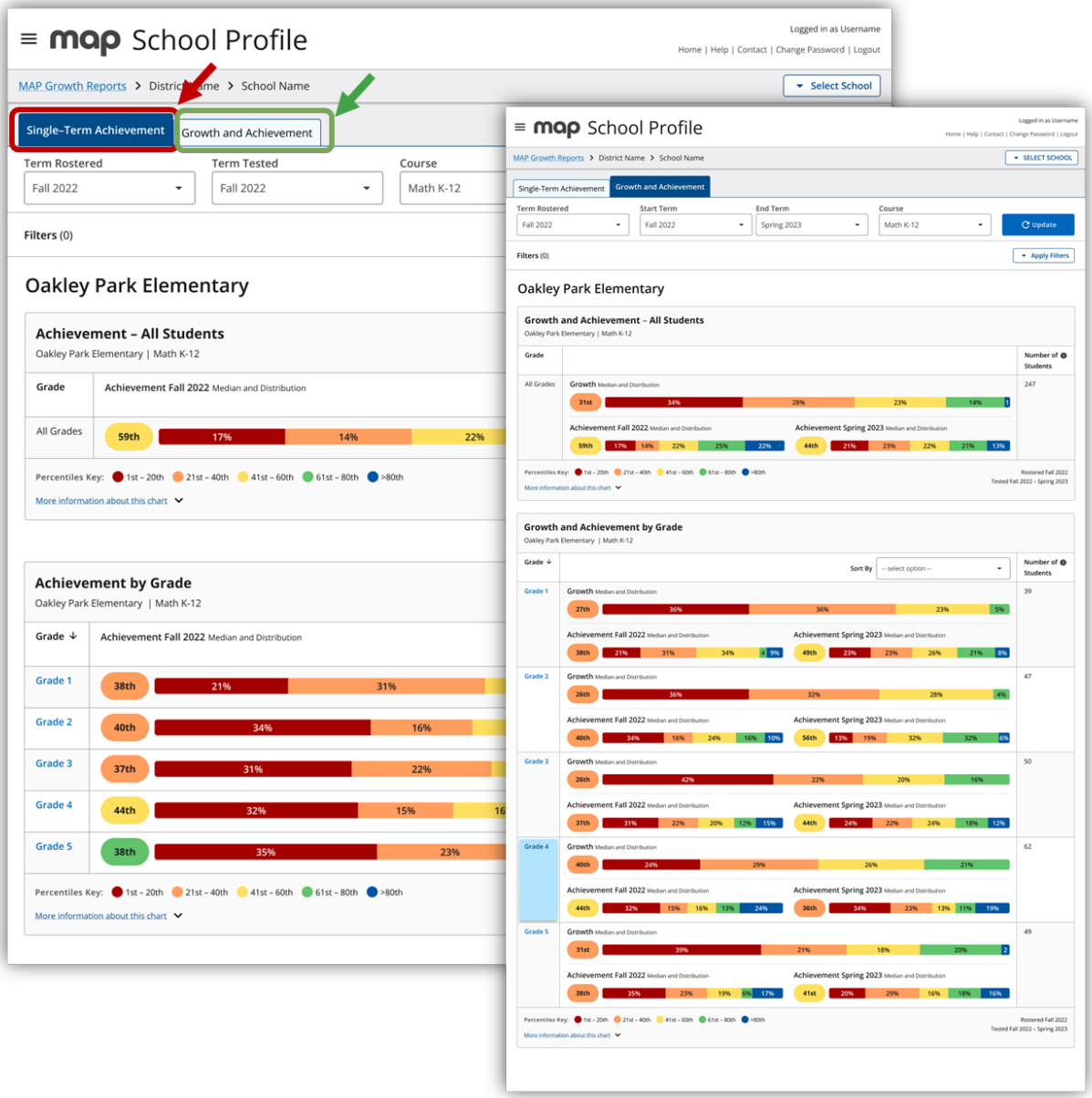
10.4.1. School Profile Report

The School Profile report is a set of refinable data views that provides school and district leaders with a rich and focused experience using their school's MAP Growth data.

Term-based growth medians and aggregate data, sorted by school and grade, illuminate a school's baseline, while changes in student data over time help leaders identify which instructional practices and programs lead to improved student outcomes.

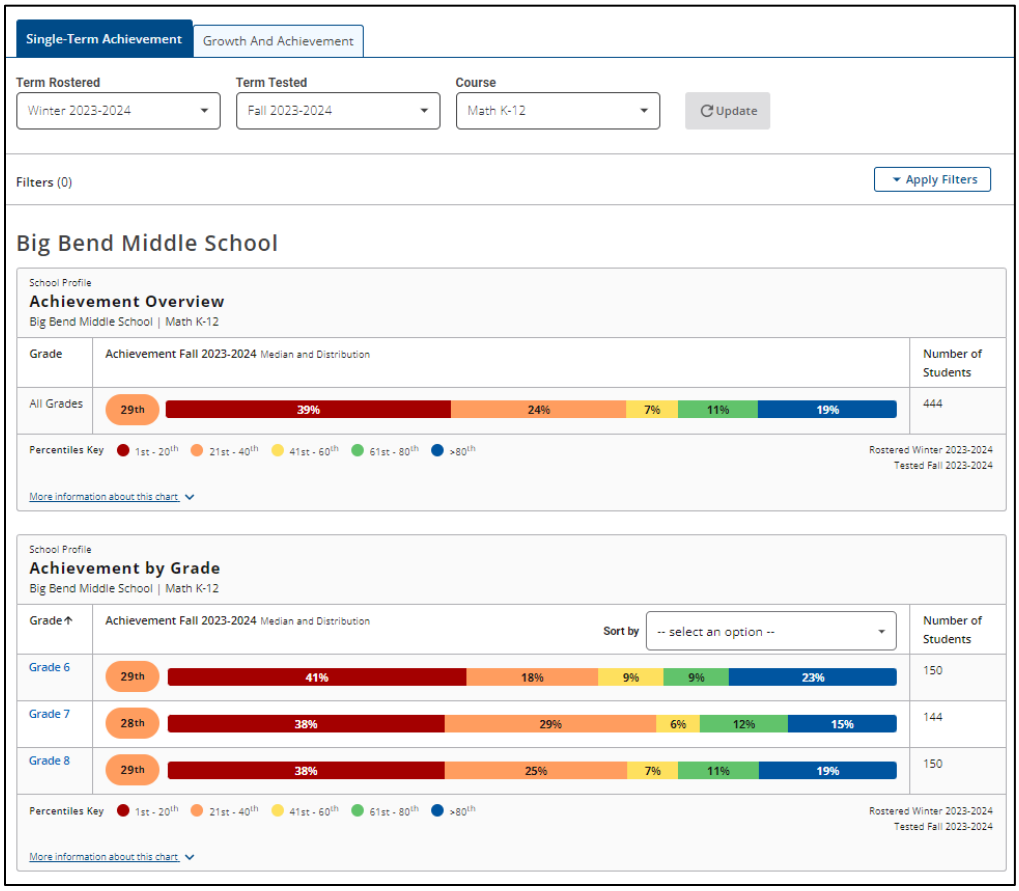
These insights—which can be filtered by ethnicity, gender, and/or program—enable school and district leaders to make quick and confident decisions that help support student growth. In addition to grade-level data, educators can drill down to view class-level data by grade. Figure 10.9 presents an example of a School Profile report.

Figure 10.9. Example School Profile Report



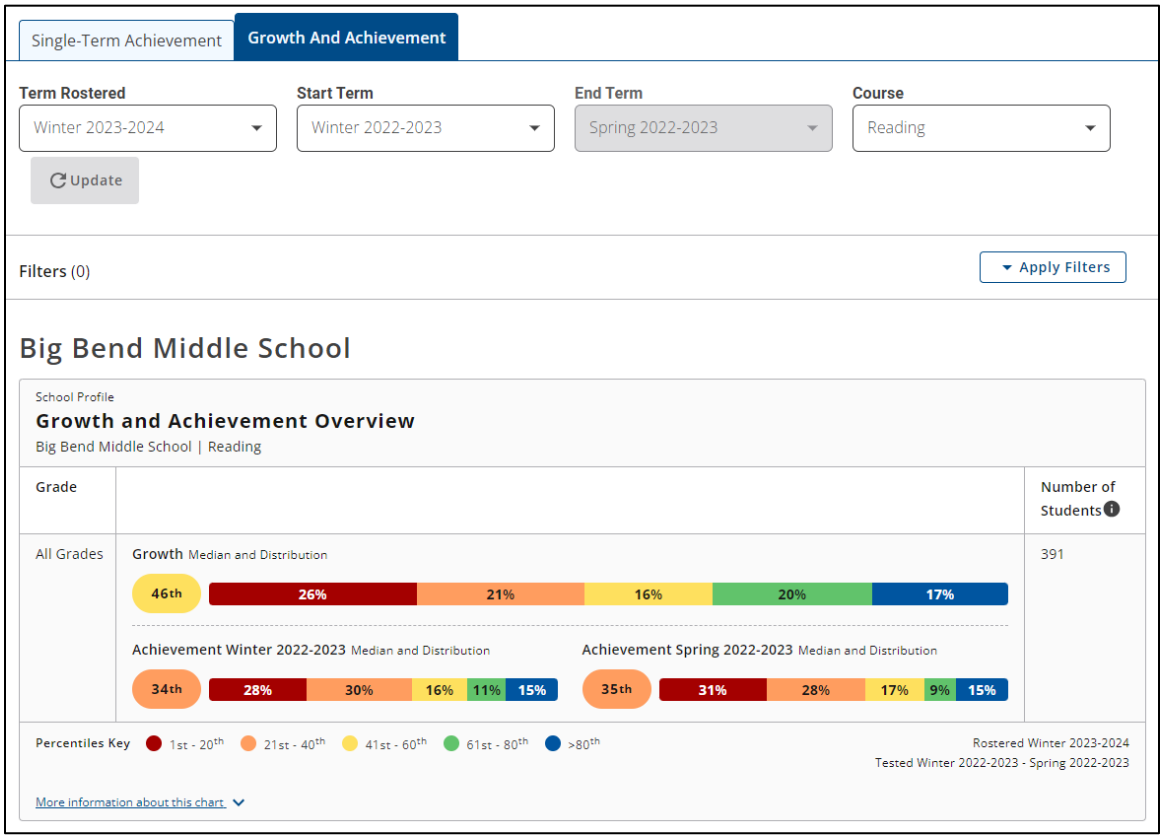
The Single-Term Achievement tab of the report, highlighted in Figure 10.10, presents single-term achievement data for all grades in a school. Educators can review the median achievement percentile as well as a breakdown of achievement percentiles by quintile. They can also see the number of students with a valid growth measure in a particular population.

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



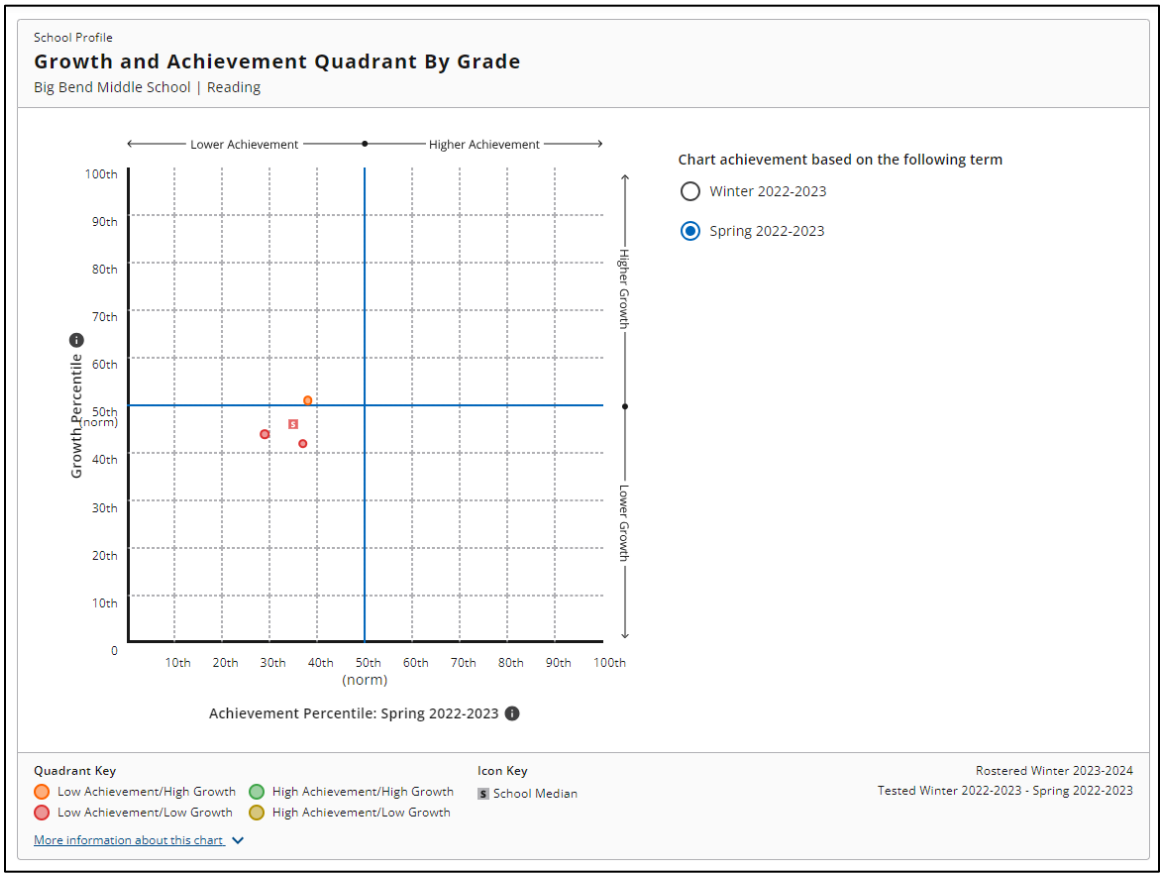
The Growth and Achievement tab of the School Profile report, highlighted in Figure 10.11, includes three modules to help educators explore data across two terms. The Growth and Achievement Overview module presents overall growth and achievement comparisons.

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



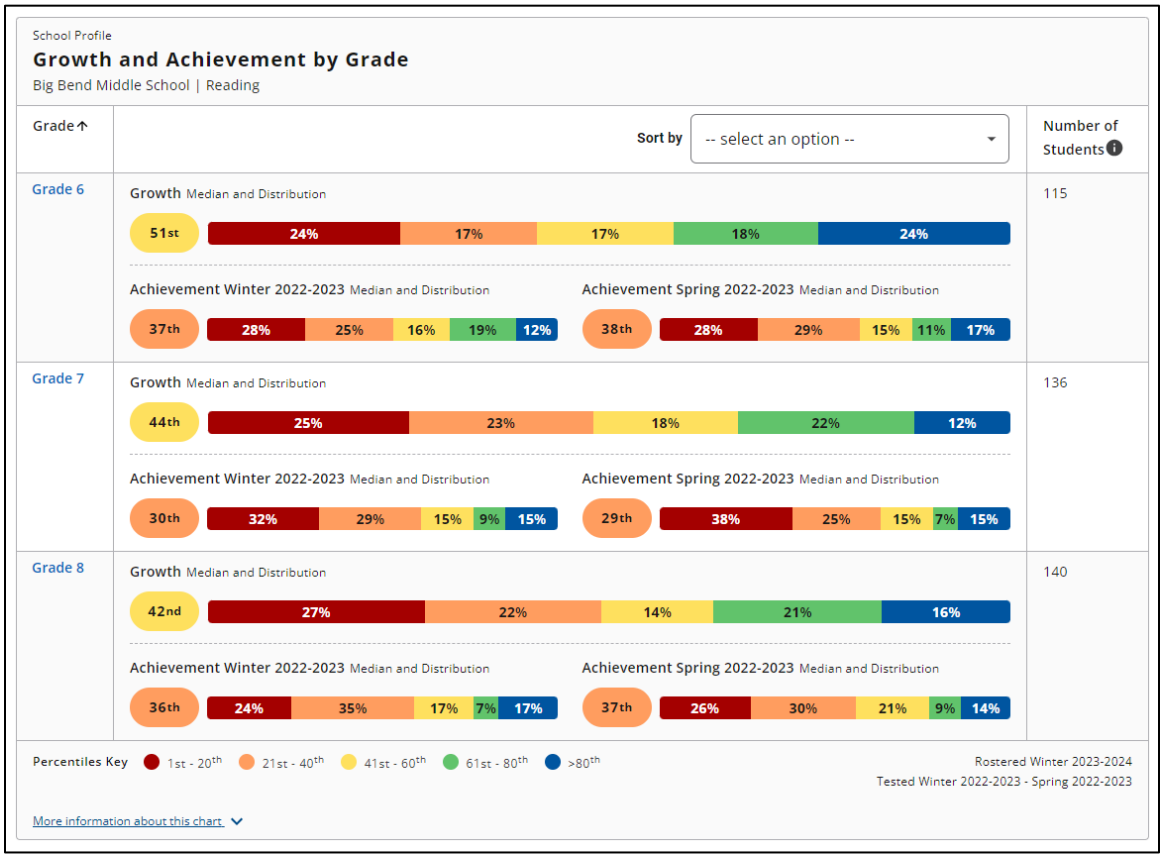
The Growth and Achievement Quadrant by Grade module shown in Figure 10.12 offers a quadrant view of growth and achievement data across two terms. The *x*-axis position shows the grade's median achievement percentile, which is a measure of how the grade performed in a single term. The *y*-axis position shows the grade's median growth percentile. This is a measure of the grade's growth between the start term and end term.

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 and enables educators to select a grade for class-level information or select a percentile band to view student 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 assess performance trends by grade and school. This report enables district administrators to monitor student performance and growth over time to support critical decision-making about when and how to invest in programs, interventions, instructional supports, and curricular tools.

The District Profile provides various levels of aggregated achievement and growth data:

- Total district aggregate: All students in all grades in all schools within a district
- Grade-level data: All students in a single grade across all schools within a district

Like the School Profile, the District Profile includes tabs for Single-Term Achievement (Figure 10.14) and Growth and Achievement (Figure 10.15). Data can be filtered by ethnicity, gender, and/or program. Educators can select a grade to see schools compared within that grade level.

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

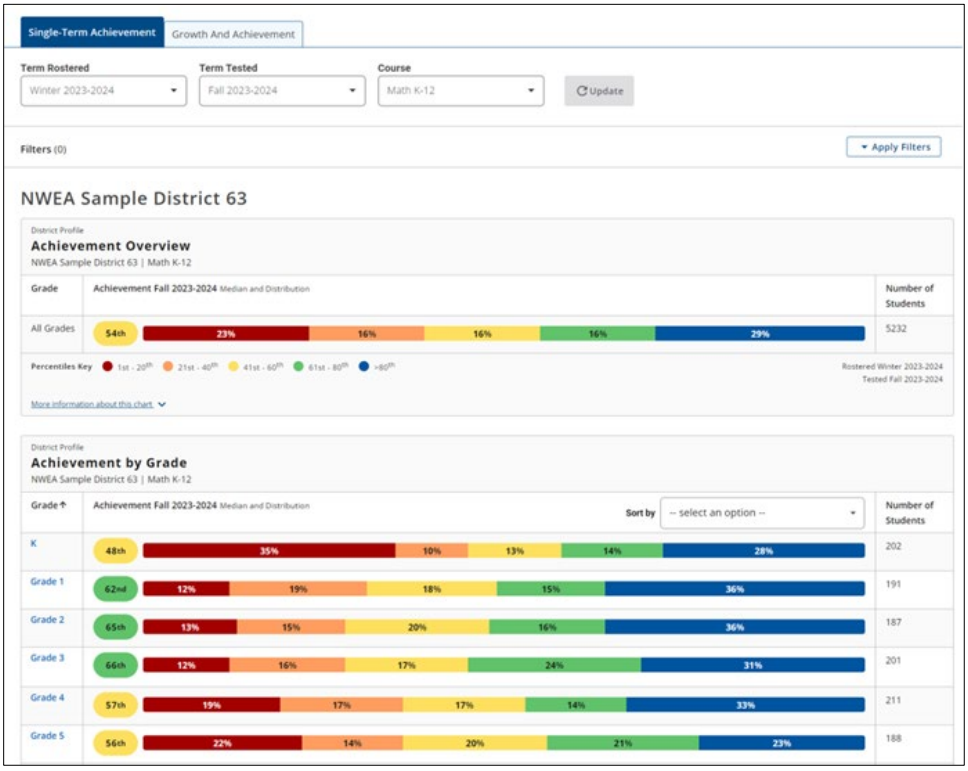
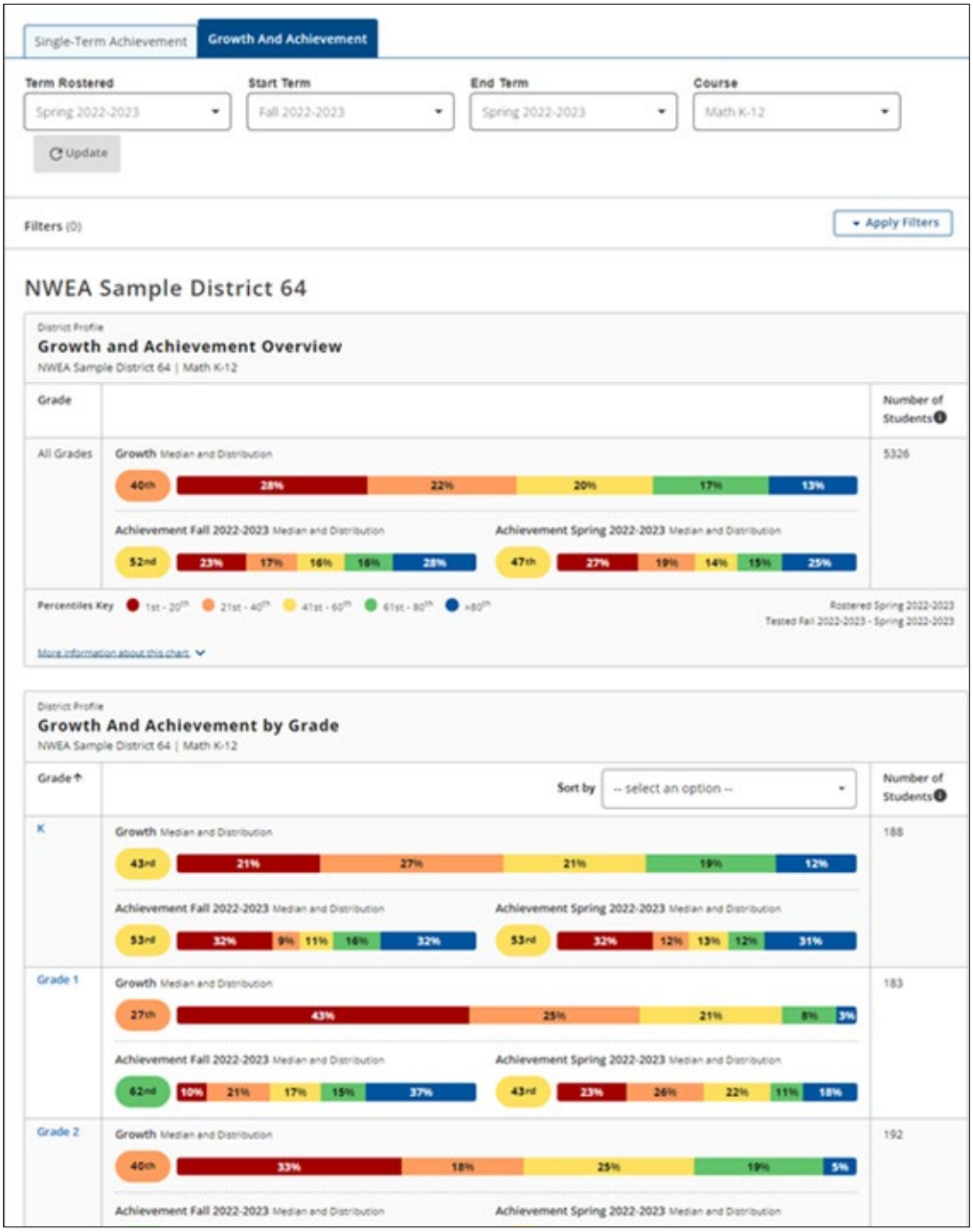


Figure 10.15. Example of Growth and Achievement Tab in District Profile Report



10.5. Score Interpretation Guide

The *MAP Growth Reports Portfolio*, our score interpretation guide for MAP Growth, includes a description and sample of each report. It is accessible anytime by districts and schools online at https://www.nwea.org/resource-center/resource/map-growth-reports-portfolio.

10.6. Professional Learning

No matter what level of expertise educators and administrators have with assessments and data interpretation, NWEA professional learning helps them become more confident and comfortable accessing, understanding, using, and sharing data. Whether in-person, virtual, or asynchronous

learning online, NWEA provides educators with the tools and training they need to inform instructional decision-making and enhance learning outcomes for all students.

NWEA professional learning is evidence-based, grounded in the work of leading researchers, and aligned with the Learning Forward Standards for Professional Learning (SPL), the Interstate New Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards, and the Every Student Succeeds Act. It consistently receives a 93% or higher customer experience rating from workshop participants.

Table 10.1 provides a sample professional learning plan to support initial administration and data analysis. Each school district and learning agency is unique, and NWEA works with school leadership to build a comprehensive professional learning plan that supports long-term success. The workshops described in the table are part of the MAP Growth Learning Pathway, a series that empowers educators to feel confident, connected, and equipped to use MAP Growth to drive student success.

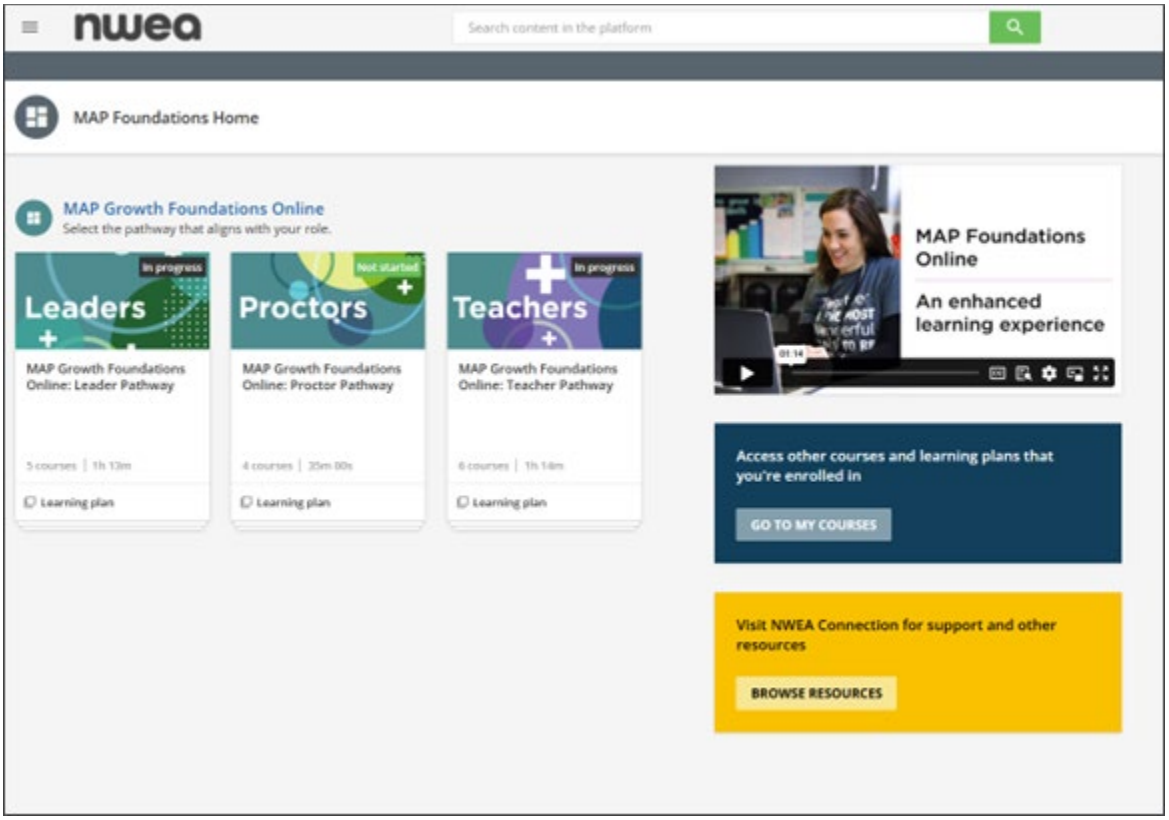
Table 10.1. Professional Learning Workshops

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 have 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 a support-focused online resource center that provides educators and education leaders with a place to find information and ask questions about NWEA products and services, including live chat support. NWEA Connection is part of a comprehensive approach to support and learning.

On NWEA Connection, educators can:

- Chat with an NWEA Partner Support representative
- Access content targeted to support successful assessment use
- Quickly search the Support Knowledge Base for answers to product questions
- See what is changing with NWEA products
- View their 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 to support educators and administrators on a wide range of topics. Webinars, eBooks, white papers, case studies, guides, briefs, and more are available and easily accessible. Also available on the website is the *Teach. Learn. Grow.* blog, with tips and discussions on topics important to educators.

10.8.4. NWEA YouTube Channel

Explore short videos, deep-dive resources, recorded webinars, and curated playlists—such as "MAP Growth 101" and "Norms 101"—all in one place. Explore what is available on our channel at 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

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

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 Erbaum 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

He, W., & Meyer, P. (2021). *MAP Growth universal screening benchmarks: Establishing MAP Growth as an effective universal screener*. NWEA Research Report. NWEA. https://www.nwea.org/uploads/MAP-Growth-Universal-Screening-Benchmarks-2021-03-12_NWEA_report-with-Memo.pdf

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

National Center on Intensive Intervention (NCII). (2020). *Academic screening tools chart rating rubrics*. https://intensiveintervention.org/sites/default/files/2025-02/NCII_AS_RatingRubric_2025_508.pdf

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 (2025). *2025 MAP Growth norms technical manual*. NWEA Research Report. NWEA. https://www.nwea.org/resource-center/white-paper/88182/MAP-Growth-Norms_NWEA_Technical-Manual.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

Thum, Y. M. (2017). *MAP Growth college readiness benchmarks: An addendum with preliminary results keyed on the SAT*. NWEA Research Report. NWEA. https://www.nwea.org/uploads/2020/10/MAP-SAT-College-Readiness-Benchmarks_NWEA_linkingstudy.pdf

Thum, Y. M., & Matta, T. (2015). *MAP college readiness benchmarks: A research brief*. NWEA Research Report. NWEA. https://www.nwea.org/content/uploads/2015/08/MAP-College-Readiness-Benchmark_Study-AUG15-Revised.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 Subject, State, and Term

Table A1. MAP Growth Reliability Results

Subject	State	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Math	AK	60,155	0.98	53,858	0.98	53,611	0.98
Math	AL	19,400	0.99	19,805	0.98	19,099	0.99
Math	AR	45,663	0.99	34,837	0.99	34,418	0.99
Math	AZ	147,419	0.99	144,071	0.99	143,038	0.99
Math	CA	632,235	0.99	553,216	0.99	536,103	0.99
Math	CO	261,115	0.99	238,035	0.99	245,977	0.99
Math	CT	71,376	0.99	64,813	0.99	58,310	0.99
Math	DC	36,841	0.99	31,301	0.99	34,608	0.98
Math	DE	22,826	0.99	22,601	0.99	21,983	0.99
Math	FL	177,966	0.99	161,402	0.99	153,121	0.99
Math	GA	494,005	0.99	474,261	0.99	430,413	0.99
Math	HI	12,247	0.98	5,631	0.98	7,912	0.98
Math	IA	40,108	0.99	33,470	0.98	28,890	0.98
Math	ID	18,586	0.99	14,046	0.99	15,301	0.99
Math	IL	502,235	0.99	471,733	0.99	497,014	0.99
Math	IN	314,050	0.99	291,592	0.99	256,859	0.99
Math	KS	45,162	0.99	45,481	0.99	42,406	0.98
Math	KY	240,375	0.99	219,817	0.99	227,752	0.99
Math	LA	39,849	0.99	37,796	0.99	36,394	0.99
Math	MA	89,764	0.99	82,225	0.99	83,525	0.99
Math	MD	53,574	0.98	104,124	0.98	48,714	0.98
Math	ME	30,308	0.99	24,256	0.99	79,315	0.98
Math	MI	678,252	0.99	566,975	0.99	663,888	0.99
Math	MN	100,676	0.99	59,830	0.99	78,631	0.99
Math	MO	146,421	0.99	148,601	0.99	145,694	0.99
Math	MS	27,259	0.99	15,901	0.98	15,844	0.98
Math	MT	39,633	0.99	28,061	0.99	38,282	0.99
Math	NC	139,118	0.99	128,430	0.99	132,077	0.99
Math	ND	21,496	0.99	17,424	0.98	20,269	0.98
Math	NE	169,003	0.99	152,172	0.99	100,683	0.97
Math	NH	26,497	0.99	20,806	0.98	19,394	0.99
Math	NJ	142,680	0.99	122,911	0.99	142,162	0.99
Math	NM	40,515	0.99	38,819	0.98	39,257	0.98
Math	NV	277,200	0.99	268,967	0.99	277,585	0.99
Math	NY	446,366	0.99	399,245	0.99	429,779	0.99
Math	OH	479,447	0.99	459,630	0.99	440,168	0.99
Math	OK	127,985	0.99	125,537	0.99	121,006	0.99
Math	OR	82,943	0.99	70,425	0.99	56,575	0.99
Math	PA	183,010	0.99	179,517	0.99	183,714	0.99
Math	SC	201,869	0.99	197,777	0.99	202,574	0.99

Subject	State	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Math	SD	69,570	0.99	67,859	0.99	63,505	0.99
Math	TN	79,769	0.99	73,892	0.99	75,056	0.99
Math	TX	1,904,861	0.99	1,816,441	0.99	1,885,934	0.99
Math	UT	38,465	0.99	33,769	0.99	36,288	0.99
Math	VA	212,212	0.98	195,868	0.98	200,986	0.98
Math	VT	7,926	0.99	7,959	0.99	6,378	0.98
Math	WA	112,224	0.99	78,557	0.99	111,726	0.99
Math	WI	111,683	0.99	93,469	0.99	110,265	0.99
Reading	AK	58,159	0.98	51,911	0.98	53,400	0.95
Reading	AL	25,627	0.99	22,848	0.98	22,801	0.98
Reading	AR	48,259	0.99	39,441	0.98	36,089	0.98
Reading	AZ	154,115	0.98	151,789	0.98	152,336	0.98
Reading	CA	662,743	0.98	550,235	0.98	527,779	0.98
Reading	CO	251,757	0.98	227,340	0.98	238,645	0.98
Reading	CT	56,717	0.98	49,497	0.98	44,463	0.98
Reading	DC	48,749	0.98	43,666	0.98	45,159	0.98
Reading	DE	23,032	0.99	23,046	0.98	22,432	0.98
Reading	FL	292,588	0.98	249,292	0.98	224,587	0.98
Reading	GA	578,646	0.99	549,214	0.98	502,118	0.98
Reading	HI	12,107	0.98	5,314	0.98	7,729	0.98
Reading	IA	49,530	0.98	44,245	0.98	30,994	0.98
Reading	ID	22,545	0.98	17,517	0.98	16,839	0.98
Reading	IL	500,399	0.99	469,430	0.98	491,175	0.98
Reading	IN	316,820	0.99	294,931	0.99	253,709	0.98
Reading	KS	50,983	0.99	46,038	0.98	47,557	0.98
Reading	KY	239,727	0.99	219,846	0.99	228,403	0.98
Reading	LA	49,491	0.99	45,621	0.98	40,093	0.98
Reading	MA	89,418	0.99	82,188	0.98	82,651	0.98
Reading	MD	94,791	0.98	117,299	0.98	64,082	0.98
Reading	ME	29,976	0.99	22,849	0.99	78,780	0.96
Reading	MI	675,817	0.99	562,595	0.98	661,716	0.98
Reading	MN	96,879	0.98	58,009	0.98	75,711	0.98
Reading	MO	159,459	0.98	156,044	0.98	148,728	0.98
Reading	MS	26,922	0.99	16,355	0.98	15,214	0.98
Reading	MT	40,234	0.99	28,112	0.98	37,874	0.98
Reading	NC	106,304	0.98	106,397	0.98	107,833	0.98
Reading	ND	22,159	0.98	17,979	0.98	20,262	0.98
Reading	NE	164,851	0.99	147,903	0.98	98,917	0.94
Reading	NH	26,681	0.98	20,738	0.98	19,853	0.98
Reading	NJ	157,477	0.98	130,445	0.98	155,086	0.98
Reading	NM	43,311	0.98	40,829	0.98	41,148	0.98
Reading	NV	307,442	0.99	310,195	0.99	312,561	0.98

Subject	State	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Reading	NY	497,118	0.99	446,187	0.98	468,647	0.98
Reading	OH	521,948	0.99	498,312	0.98	459,093	0.98
Reading	OK	128,450	0.99	126,821	0.99	120,903	0.98
Reading	OR	77,291	0.98	62,842	0.98	49,586	0.98
Reading	PA	200,504	0.98	187,021	0.98	196,323	0.98
Reading	SC	209,041	0.99	203,776	0.98	206,244	0.98
Reading	SD	65,465	0.99	63,987	0.98	59,040	0.98
Reading	TN	49,462	0.98	44,132	0.98	45,541	0.98
Reading	TX	1,957,220	0.99	1,818,164	0.98	1,914,787	0.98
Reading	UT	68,762	0.98	66,694	0.98	67,549	0.98
Reading	VA	247,054	0.98	219,154	0.98	217,057	0.98
Reading	VT	7,640	0.98	7,885	0.98	6,012	0.98
Reading	WA	108,636	0.98	79,597	0.98	106,768	0.98
Reading	WI	103,914	0.98	87,218	0.98	100,045	0.98
Language	AL	7,689	0.97	6,866	0.97	6,252	0.97
Language	AR	9,163	0.97	8,594	0.97	9,198	0.96
Language	AZ	54,539	0.97	52,344	0.97	50,646	0.97
Language	CA	142,181	0.97	100,702	0.97	107,851	0.97
Language	CO	53,074	0.97	40,686	0.97	47,460	0.97
Language	CT	9,908	0.96	8,697	0.96	9,429	0.96
Language	DC	10,001	0.97	6,456	0.97	10,123	0.96
Language	FL	67,496	0.97	56,955	0.97	70,049	0.97
Language	GA	150,982	0.98	142,097	0.98	141,363	0.98
Language	IA	10,434	0.97	5,855	0.96	5,267	0.96
Language	ID	9,558	0.97	6,862	0.96	7,791	0.96
Language	IL	47,371	0.97	35,243	0.97	42,670	0.96
Language	IN	63,462	0.98	50,372	0.97	51,032	0.97
Language	KS	10,705	0.97	8,060	0.97	9,913	0.97
Language	KY	58,946	0.97	51,722	0.97	54,440	0.97
Language	MA	19,412	0.97	14,375	0.96	18,652	0.96
Language	MD	16,679	0.97	13,632	0.97	16,567	0.96
Language	MI	109,898	0.97	76,172	0.97	100,686	0.96
Language	MN	19,705	0.97	7,758	0.96	18,839	0.96
Language	MO	41,761	0.97	46,919	0.97	40,257	0.96
Language	MT	14,732	0.97	7,458	0.97	13,438	0.96
Language	NC	12,217	0.97	9,350	0.97	11,748	0.96
Language	NE	36,465	0.97	17,469	0.96	24,750	0.96
Language	NJ	51,031	0.98	33,438	0.98	50,971	0.98
Language	NM	17,050	0.98	15,655	0.97	15,913	0.97
Language	NV	17,484	0.98	15,035	0.97	22,864	0.97
Language	NY	23,309	0.97	20,370	0.97	21,420	0.97
Language	OH	46,460	0.97	35,498	0.97	41,437	0.96

Subject	State	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Language	OK	25,987	0.97	24,322	0.97	22,758	0.97
Language	OR	18,024	0.98	9,269	0.97	18,100	0.97
Language	PA	28,741	0.97	22,300	0.97	25,350	0.97
Language	SC	18,538	0.97	9,738	0.97	11,353	0.97
Language	SD	16,690	0.97	15,639	0.97	15,186	0.97
Language	TN	10,515	0.97	9,763	0.97	9,714	0.97
Language	TX	194,892	0.98	177,073	0.97	186,989	0.97
Language	UT	9,386	0.97	7,893	0.98	5,351	0.97
Language	VA	16,481	0.97	15,805	0.96	17,304	0.96
Language	WA	22,231	0.97	11,926	0.96	21,463	0.96
Language	WI	22,527	0.97	18,389	0.96	22,009	0.96
Science	AR	11,893	0.95	9,210	0.95	12,382	0.95
Science	AZ	22,921	0.95	22,931	0.95	21,834	0.95
Science	CA	81,965	0.95	70,716	0.95	58,563	0.95
Science	CO	52,658	0.95	45,856	0.95	50,229	0.95
Science	FL	76,743	0.95	76,413	0.95	60,929	0.94
Science	GA	73,902	0.96	73,389	0.96	73,969	0.96
Science	IA	17,221	0.94	13,353	0.94	11,657	0.94
Science	IL	36,257	0.95	31,869	0.95	39,257	0.95
Science	IN	8,943	0.94	7,566	0.95	6,732	0.95
Science	KY	39,055	0.95	32,946	0.95	32,862	0.94
Science	LA	8,448	0.95	8,463	0.96	7,944	0.96
Science	MA	15,983	0.95	13,795	0.95	13,218	0.95
Science	MI	92,235	0.95	67,402	0.95	84,289	0.95
Science	MO	42,963	0.95	39,506	0.95	35,301	0.94
Science	NC	7,350	0.95	6,682	0.95	6,139	0.95
Science	NE	62,942	0.96	24,756	0.94	44,158	0.96
Science	NJ	32,802	0.95	27,577	0.95	33,849	0.96
Science	NM	15,107	0.95	14,800	0.95	14,449	0.96
Science	NV	14,580	0.95	13,130	0.94	13,154	0.94
Science	NY	24,705	0.95	23,592	0.95	23,289	0.95
Science	OH	97,389	0.95	85,849	0.95	83,755	0.95
Science	OK	18,003	0.94	17,501	0.94	16,017	0.94
Science	OR	10,132	0.96	6,181	0.96	9,170	0.96
Science	PA	23,972	0.96	23,465	0.96	23,356	0.97
Science	SD	19,979	0.95	19,254	0.95	17,831	0.95
Science	TX	821,123	0.96	703,918	0.96	826,163	0.96
Science	UT	9,237	0.96	8,049	0.96	8,301	0.96
Science	VA	8,485	0.94	7,537	0.94	8,428	0.95
Science	WA	13,381	0.94	8,069	0.94	12,388	0.94
Science	WI	7,893	0.95	6,136	0.95	7,961	0.94

Note. Results were omitted for a state whenever the sample size was less than 5,000 for any term. Language = Language Usage

Appendix B: Examinee Demographics by Subject, Grade, and Term

Table B1. MAP Growth Demographics for Gender

Subject	Grade	Term	Total	Percentage Male	Percentage Female
Math	K	Fall	613,477	50.37	49.63
Math	K	Winter	656,170	50.43	49.57
Math	K	Spring	695,210	50.44	49.56
Math	1	Fall	824,366	50.49	49.51
Math	1	Winter	795,374	50.52	49.48
Math	1	Spring	833,368	50.46	49.54
Math	2	Fall	934,036	50.66	49.34
Math	2	Winter	902,893	50.65	49.35
Math	2	Spring	936,343	50.62	49.38
Math	3	Fall	997,603	50.68	49.32
Math	3	Winter	950,854	50.69	49.31
Math	3	Spring	937,002	50.62	49.38
Math	4	Fall	973,686	50.71	49.29
Math	4	Winter	922,418	50.73	49.27
Math	4	Spring	904,291	50.68	49.32
Math	5	Fall	982,966	50.88	49.12
Math	5	Winter	931,722	50.90	49.10
Math	5	Spring	910,155	50.88	49.12
Math	6	Fall	1,004,485	50.89	49.11
Math	6	Winter	909,742	50.92	49.08
Math	6	Spring	908,214	50.88	49.12
Math	7	Fall	1,001,808	50.92	49.08
Math	7	Winter	885,406	50.95	49.05
Math	7	Spring	891,234	50.92	49.08
Math	8	Fall	904,405	50.92	49.08
Math	8	Winter	794,292	51.01	48.99
Math	8	Spring	775,040	50.92	49.08
Math	9	Fall	390,063	51.56	48.44
Math	9	Winter	297,870	51.66	48.34
Math	9	Spring	317,181	51.49	48.51
Math	10	Fall	310,924	51.51	48.49
Math	10	Winter	239,622	51.64	48.36
Math	10	Spring	256,550	51.32	48.68
Math	11	Fall	192,821	51.60	48.40
Math	11	Winter	146,942	51.99	48.01
Math	11	Spring	139,749	51.64	48.36
Math	12	Fall	97,487	52.18	47.82
Math	12	Winter	69,124	52.47	47.53
Math	12	Spring	53,090	52.12	47.88
Reading	K	Fall	483,778	50.34	49.66
Reading	K	Winter	531,151	50.43	49.57

Subject	Grade	Term	Total	Percentage Male	Percentage Female
Reading	K	Spring	568,438	50.43	49.57
Reading	1	Fall	681,972	50.48	49.52
Reading	1	Winter	661,080	50.53	49.47
Reading	1	Spring	694,784	50.48	49.52
Reading	2	Fall	821,485	50.65	49.35
Reading	2	Winter	797,706	50.67	49.33
Reading	2	Spring	840,357	50.61	49.39
Reading	3	Fall	969,334	50.66	49.34
Reading	3	Winter	923,215	50.69	49.31
Reading	3	Spring	914,571	50.62	49.38
Reading	4	Fall	963,626	50.73	49.27
Reading	4	Winter	901,749	50.75	49.25
Reading	4	Spring	884,014	50.69	49.31
Reading	5	Fall	975,673	50.88	49.12
Reading	5	Winter	916,626	50.92	49.08
Reading	5	Spring	897,601	50.88	49.12
Reading	6	Fall	1,042,782	50.98	49.02
Reading	6	Winter	942,857	51.02	48.98
Reading	6	Spring	932,747	50.95	49.05
Reading	7	Fall	1,048,027	51.14	48.86
Reading	7	Winter	925,648	51.25	48.75
Reading	7	Spring	915,331	51.17	48.83
Reading	8	Fall	1,033,172	51.19	48.81
Reading	8	Winter	911,102	51.30	48.70
Reading	8	Spring	882,124	51.20	48.80
Reading	9	Fall	678,675	51.75	48.25
Reading	9	Winter	542,957	51.88	48.12
Reading	9	Spring	547,900	51.74	48.26
Reading	10	Fall	560,256	51.60	48.40
Reading	10	Winter	439,925	51.77	48.23
Reading	10	Spring	448,724	51.51	48.49
Reading	11	Fall	308,456	52.41	47.59
Reading	11	Winter	229,239	52.82	47.18
Reading	11	Spring	211,248	52.45	47.55
Reading	12	Fall	169,834	52.72	47.28
Reading	12	Winter	117,531	53.05	46.95
Reading	12	Spring	95,831	52.74	47.26
Language	2	Fall	105,861	50.06	49.94
Language	2	Winter	94,991	50.07	49.93
Language	2	Spring	109,142	49.98	50.02
Language	3	Fall	179,643	50.02	49.98
Language	3	Winter	155,066	50	50
Language	3	Spring	169,881	50	50.01

Subject	Grade	Term	Total	Percentage Male	Percentage Female
Language	4	Fall	181,731	50.05	49.95
Language	4	Winter	157,660	50.03	49.97
Language	4	Spring	170,090	50.03	49.97
Language	5	Fall	188,766	50.35	49.65
Language	5	Winter	163,313	50.46	49.54
Language	5	Spring	178,188	50.29	49.71
Language	6	Fall	204,477	50.57	49.43
Language	6	Winter	166,715	50.64	49.36
Language	6	Spring	187,054	50.57	49.43
Language	7	Fall	199,765	50.75	49.25
Language	7	Winter	159,544	50.83	49.17
Language	7	Spring	181,344	50.70	49.30
Language	8	Fall	196,607	50.97	49.03
Language	8	Winter	156,288	51	49
Language	8	Spring	173,749	50.88	49.12
Language	9	Fall	113,737	51.87	48.13
Language	9	Winter	74,472	52.17	47.83
Language	9	Spring	90,056	52.04	47.96
Language	10	Fall	97,977	51.83	48.17
Language	10	Winter	65,633	52.01	47.99
Language	10	Spring	78,526	51.89	48.11
Language	11	Fall	66,487	52.18	47.82
Language	11	Winter	43,837	52.67	47.33
Language	11	Spring	46,124	52.35	47.65
Language	12	Fall	37,362	52.55	47.45
Language	12	Winter	24,423	52.67	47.33
Language	12	Spring	20,499	52.72	47.28
Science	2	Fall	37,898	51.08	48.92
Science	2	Winter	36,353	51.02	48.98
Science	2	Spring	39,377	50.97	49.03
Science	3	Fall	159,068	50.43	49.57
Science	3	Winter	135,125	50.57	49.43
Science	3	Spring	157,732	50.39	49.61
Science	4	Fall	218,812	50.45	49.55
Science	4	Winter	181,242	50.53	49.47
Science	4	Spring	220,728	50.50	49.50
Science	5	Fall	318,188	50.67	49.33
Science	5	Winter	280,182	50.77	49.23
Science	5	Spring	300,225	50.66	49.34
Science	6	Fall	288,517	50.75	49.25
Science	6	Winter	242,384	50.81	49.19
Science	6	Spring	275,991	50.71	49.29
Science	7	Fall	304,931	50.89	49.12

Subject	Grade	Term	Total	Percentage Male	Percentage Female
Science	7	Winter	258,665	51.03	48.97
Science	7	Spring	294,714	50.86	49.14
Science	8	Fall	336,460	50.97	49.03
Science	8	Winter	287,102	51.08	48.92
Science	8	Spring	301,412	50.91	49.09
Science	9	Fall	60,686	51.34	48.66
Science	9	Winter	47,129	51.38	48.62
Science	9	Spring	52,219	51.22	48.78
Science	10	Fall	56,559	51.33	48.67
Science	10	Winter	45,128	51.22	48.78
Science	10	Spring	45,846	51.45	48.55
Science	11	Fall	39,046	51.50	48.50
Science	11	Winter	30,351	51.84	48.16
Science	11	Spring	26,318	52.10	47.90
Science	12	Fall	15,643	51.67	48.33
Science	12	Winter	11,399	52.29	47.71
Science	12	Spring	9,088	52.15	47.85

Note. Language = Language Usage.

Table B.2. MAP Growth Demographics for Race/Ethnicity

Subject	Grade	Term	Total	Percentage							
				Asian	Black	NH/PI	AI/AN	Hispanic	White	Other	Multiple
Math	K	Fall	613,477	4.91	15.77	0.24	1.43	26.29	40.17	5.96	5.22
Math	K	Winter	656,170	4.87	15.61	0.25	1.37	26.20	40.70	5.80	5.22
Math	K	Spring	695,210	5	15.65	0.25	1.33	25.62	40.90	5.78	5.49
Math	1	Fall	824,366	5.11	15.61	0.23	1.28	24.98	41.59	5.78	5.42
Math	1	Winter	795,374	5.08	15.69	0.23	1.29	25.59	41.25	5.52	5.35
Math	1	Spring	833,368	5.18	15.63	0.23	1.24	25.03	41.24	5.70	5.77
Math	2	Fall	934,036	5.29	14.95	0.23	1.27	24.69	42.22	6.03	5.32
Math	2	Winter	902,893	5.22	15.33	0.23	1.25	25.31	41.76	5.61	5.29
Math	2	Spring	936,343	5.34	15.03	0.23	1.22	24.78	41.79	5.99	5.61
Math	3	Fall	997,603	5.42	14.66	0.25	1.33	24.59	42.34	6.03	5.39
Math	3	Winter	950,854	5.42	15.01	0.25	1.30	25.25	41.98	5.43	5.35
Math	3	Spring	937,002	5.60	14.98	0.23	1.21	24.94	41.62	5.79	5.63
Math	4	Fall	973,686	5.46	14.67	0.26	1.30	24.17	42.82	6.05	5.26
Math	4	Winter	922,418	5.40	14.94	0.26	1.30	24.89	42.47	5.55	5.19
Math	4	Spring	904,291	5.55	14.77	0.23	1.19	24.44	42.28	6.02	5.53
Math	5	Fall	982,966	5.46	14.53	0.27	1.33	23.92	43.20	6.16	5.13
Math	5	Winter	931,722	5.44	14.95	0.27	1.34	24.61	42.83	5.52	5.05
Math	5	Spring	910,155	5.62	14.85	0.24	1.21	24.22	42.45	6.06	5.35
Math	6	Fall	1,004,485	5.61	14.52	0.28	1.30	23.69	42.84	6.70	5.05
Math	6	Winter	909,742	5.71	15.19	0.28	1.35	24.20	42.68	5.55	5.04
Math	6	Spring	908,214	5.90	14.76	0.26	1.18	23.87	42.16	6.53	5.33
Math	7	Fall	1,001,808	5.55	14.70	0.28	1.32	24.06	42.66	6.50	4.93
Math	7	Winter	885,406	5.56	15.41	0.29	1.35	24.90	42.31	5.33	4.85
Math	7	Spring	891,234	5.82	14.79	0.26	1.18	24.35	41.91	6.54	5.14
Math	8	Fall	904,405	4.79	15.27	0.27	1.39	24.19	42.56	6.63	4.89
Math	8	Winter	794,292	4.84	16.23	0.27	1.44	25.09	42.11	5.17	4.84
Math	8	Spring	775,040	4.91	15.74	0.26	1.29	24.75	41.46	6.50	5.10
Math	9	Fall	390,063	4.65	15.19	0.47	2.05	28.85	37.47	7.12	4.21
Math	9	Winter	297,870	5.12	16.41	0.46	2.08	30.34	36.08	5.09	4.42

Subject	Grade	Term	Total	Percentage							
				Asian	Black	NH/PI	AI/AN	Hispanic	White	Other	Multiple
Math	9	Spring	317,181	4.38	14.28	0.44	1.87	29.36	37.39	7.87	4.42
Math	10	Fall	310,924	4.26	14.27	0.47	2.24	30.69	36.80	7.51	3.75
Math	10	Winter	239,622	4.57	15.10	0.46	2.32	31.71	36.38	5.43	4.03
Math	10	Spring	256,550	4.28	14.05	0.47	2.18	30.81	36.19	7.89	4.13
Math	11	Fall	192,821	4.23	14.33	0.39	2.45	31.34	35.34	8.56	3.36
Math	11	Winter	146,942	4.56	14.72	0.36	2.46	33.75	34.15	6.60	3.41
Math	11	Spring	139,749	3.83	14.25	0.38	2.31	30.70	35.22	9.16	4.16
Math	12	Fall	97,487	4.60	16.70	0.45	2.67	36.07	28.05	8.16	3.31
Math	12	Winter	69,124	4.53	16.82	0.39	2.94	38.73	26.86	6.50	3.23
Math	12	Spring	53,090	4.33	15.58	0.35	2.87	39.31	25.63	7.97	3.96
Reading	K	Fall	483,778	4.34	16.15	0.24	1.41	23.58	42.22	6.59	5.46
Reading	K	Winter	531,151	4.57	15.93	0.24	1.33	23.29	42.27	7.01	5.36
Reading	K	Spring	568,438	4.79	15.88	0.24	1.32	22.86	42.37	7.02	5.51
Reading	1	Fall	681,972	4.71	15.67	0.24	1.28	22.79	43.08	6.79	5.44
Reading	1	Winter	661,080	4.68	15.88	0.23	1.27	23.31	42.72	6.50	5.41
Reading	1	Spring	694,784	4.88	15.66	0.23	1.22	22.95	42.69	6.69	5.68
Reading	2	Fall	821,485	5.22	14.89	0.23	1.23	23.02	43.28	6.90	5.23
Reading	2	Winter	797,706	5.07	15.22	0.23	1.21	23.60	42.93	6.47	5.28
Reading	2	Spring	840,357	5.29	15.16	0.23	1.18	23.16	42.72	6.77	5.49
Reading	3	Fall	969,334	5.45	14.93	0.26	1.27	23.85	42.37	6.58	5.29
Reading	3	Winter	923,215	5.41	15.26	0.26	1.25	24.40	42.14	5.99	5.28
Reading	3	Spring	914,571	5.57	15.40	0.24	1.15	24.18	41.67	6.30	5.48
Reading	4	Fall	963,626	5.46	15.10	0.26	1.25	23.50	43.12	6.07	5.24
Reading	4	Winter	901,749	5.38	15.17	0.26	1.26	24.22	42.98	5.57	5.17
Reading	4	Spring	884,014	5.54	15	0.23	1.15	23.93	42.71	6.01	5.44
Reading	5	Fall	975,673	5.44	15	0.27	1.29	23.35	43.40	6.15	5.10
Reading	5	Winter	916,626	5.40	15.16	0.29	1.30	24.14	43.32	5.40	4.99
Reading	5	Spring	897,601	5.53	15.07	0.26	1.17	23.83	42.83	6.03	5.29
Reading	6	Fall	1,042,782	5.50	15.11	0.29	1.24	23.87	42.57	6.40	5.02
Reading	6	Winter	942,857	5.61	15.76	0.29	1.29	24.41	42.34	5.27	5.03

Subject	Grade	Term	Total	Percentage							
				Asian	Black	NH/PI	AI/AN	Hispanic	White	Other	Multiple
Reading	6	Spring	932,747	5.73	15.32	0.27	1.14	24.18	41.79	6.26	5.32
Reading	7	Fall	1,048,027	5.55	15.23	0.31	1.27	24.16	42.32	6.27	4.89
Reading	7	Winter	925,648	5.66	15.95	0.32	1.28	24.95	41.84	5.15	4.86
Reading	7	Spring	915,331	5.73	15.29	0.29	1.14	24.46	41.58	6.37	5.16
Reading	8	Fall	1,033,172	5.31	15.36	0.29	1.29	24.39	42.34	6.17	4.84
Reading	8	Winter	911,102	5.50	16.20	0.30	1.32	25.20	41.75	4.93	4.80
Reading	8	Spring	882,124	5.44	15.78	0.27	1.17	24.94	41.24	6.05	5.09
Reading	9	Fall	678,675	5.60	16.97	0.37	1.56	29.93	36.03	5.15	4.39
Reading	9	Winter	542,957	5.89	18.50	0.36	1.55	30.47	34.78	3.89	4.55
Reading	9	Spring	547,900	5.53	15.95	0.35	1.44	30.61	35.53	5.63	4.96
Reading	10	Fall	560,256	5.45	16.27	0.38	1.70	31.24	35.57	5.29	4.10
Reading	10	Winter	439,925	5.61	17.18	0.37	1.65	31.58	35.41	4.07	4.12
Reading	10	Spring	448,724	5.38	15.70	0.38	1.70	32.23	34.53	5.45	4.62
Reading	11	Fall	308,456	4.69	16.81	0.35	1.87	32.86	33.89	6.02	3.50
Reading	11	Winter	229,239	4.75	17.46	0.33	1.84	34.42	32.79	4.96	3.46
Reading	11	Spring	211,248	4.09	16.34	0.34	1.88	33.27	33.12	6.64	4.31
Reading	12	Fall	169,834	4.97	17.50	0.38	2.22	36.73	29.42	5.47	3.31
Reading	12	Winter	117,531	4.93	17.69	0.35	2.33	38	28.51	4.92	3.26
Reading	12	Spring	95,831	4.70	14.70	0.36	2.52	40.20	27.65	5.38	4.48
Language	2	Fall	105,861	4.24	12.62	0.19	1.83	16.96	46.54	12.24	5.39
Language	2	Winter	94,991	4.66	13.13	0.19	1.84	17.54	45.82	11.76	5.06
Language	2	Spring	109,142	4.71	12.56	0.23	1.59	16.05	46.20	12.98	5.69
Language	3	Fall	179,643	4.32	12.68	0.18	2.16	17.71	46.08	11.90	4.98
Language	3	Winter	155,066	4.67	14.11	0.18	2.31	18.49	44.75	10.65	4.84
Language	3	Spring	169,881	4.82	13.53	0.18	1.83	17.02	45.46	12.11	5.06
Language	4	Fall	181,731	4.19	12.52	0.19	2.26	16.92	47.10	11.87	4.95
Language	4	Winter	157,660	4.63	13.73	0.18	2.40	17.56	46	10.73	4.77
Language	4	Spring	170,090	4.74	13.17	0.17	1.95	15.95	46.70	12.33	4.99
Language	5	Fall	188,766	4.15	12.74	0.20	2.30	16.15	47.57	11.93	4.95
Language	5	Winter	163,313	4.50	14.10	0.22	2.51	16.94	46.59	10.46	4.68

Subject	Grade	Term	Total	Percentage							
				Asian	Black	NH/PI	AI/AN	Hispanic	White	Other	Multiple
Language	5	Spring	178,188	4.65	13.41	0.20	1.94	15.61	47	12.23	4.96
Language	6	Fall	204,477	3.77	12.29	0.20	2.28	15.68	48.76	11.78	5.23
Language	6	Winter	166,715	4.09	13.87	0.21	2.61	16.62	47.17	10.38	5.05
Language	6	Spring	187,054	4.42	13.14	0.21	1.97	15.19	47.38	12.41	5.27
Language	7	Fall	199,765	3.73	12.09	0.24	2.23	16.03	48.37	12.20	5.12
Language	7	Winter	159,544	4.22	14.06	0.23	2.43	17.04	46.83	10.19	5
Language	7	Spring	181,344	4.47	13.09	0.23	1.93	15.54	47.27	12.25	5.22
Language	8	Fall	196,607	3.40	12.27	0.23	2.32	16.04	49.05	11.69	5
Language	8	Winter	156,288	3.87	14.19	0.23	2.52	16.66	48.02	9.59	4.91
Language	8	Spring	173,749	4.01	13.36	0.24	1.99	15.65	47.60	12.11	5.03
Language	9	Fall	113,737	2.77	12.15	0.33	3.06	26.93	39.68	10.40	4.68
Language	9	Winter	74,472	3.02	13.04	0.37	3.86	25.01	40.69	9.38	4.63
Language	9	Spring	90,056	3.05	12.82	0.27	2.69	24.97	40.36	11.10	4.73
Language	10	Fall	97,977	2.67	12.32	0.30	3.01	28.64	38.94	10.10	4.02
Language	10	Winter	65,633	2.76	13.32	0.36	3.77	25.51	40.63	9.49	4.15
Language	10	Spring	78,526	2.98	13.01	0.30	3.01	25.83	39.38	10.85	4.63
Language	11	Fall	66,487	2.53	11.27	0.39	3.19	34.17	35.08	10.40	2.98
Language	11	Winter	43,837	2.86	11.18	0.44	4.31	30.41	36.99	10.74	3.08
Language	11	Spring	46,124	2.34	11.60	0.41	3.49	31.07	35.57	11.27	4.25
Language	12	Fall	37,362	3.25	11.53	0.38	3.85	41.96	27.20	8.66	3.17
Language	12	Winter	24,423	3.89	11.54	0.32	4.88	42.01	25.91	8.62	2.84
Language	12	Spring	20,499	2.81	10.02	0.18	4.11	44.41	23.27	10.30	4.91
Science	2	Fall	37,898	4.30	16.32	0.30	1.97	44.61	26.08	1.12	5.29
Science	2	Winter	36,353	4.48	16.64	0.29	1.47	45.86	24.44	1.54	5.27
Science	2	Spring	39,377	4.35	16.21	0.28	1.15	44.99	24	1.56	7.46
Science	3	Fall	159,068	5.96	17.69	0.21	2.16	31.99	32.53	4.38	5.09
Science	3	Winter	135,125	6.21	18.07	0.20	2.03	32.83	31.57	4.15	4.95
Science	3	Spring	157,732	6.17	17.85	0.21	1.89	31.48	32.32	4.87	5.21
Science	4	Fall	218,812	6.26	17.01	0.19	1.89	30.98	34.96	4.10	4.60
Science	4	Winter	181,242	6.26	17.03	0.20	1.86	30.17	35.77	4.05	4.66

Subject	Grade	Term	Total	Percentage							
				Asian	Black	NH/PI	AI/AN	Hispanic	White	Other	Multiple
Science	4	Spring	220,728	6.45	16.89	0.19	1.80	31.13	34.51	4.16	4.87
Science	5	Fall	318,188	5.33	16.69	0.20	1.87	31.07	36	3.95	4.90
Science	5	Winter	280,182	5.28	17.03	0.19	1.85	31.34	35.88	3.72	4.70
Science	5	Spring	300,225	5.58	16.91	0.19	1.69	31.08	35.10	4.22	5.24
Science	6	Fall	288,517	5.16	15.14	0.24	1.89	28.92	38.56	5.18	4.90
Science	6	Winter	242,384	5.46	15.65	0.24	1.94	28.80	38.36	4.51	5.04
Science	6	Spring	275,991	5.49	14.97	0.24	1.69	29.16	37.62	5.63	5.19
Science	7	Fall	304,931	4.99	14.25	0.27	1.97	29.22	39.18	5.17	4.95
Science	7	Winter	258,665	5.32	14.66	0.27	2.04	29.36	38.98	4.41	4.96
Science	7	Spring	294,714	5.23	14.49	0.26	1.85	29.72	37.98	5.23	5.24
Science	8	Fall	336,460	4.66	16.13	0.25	2	29.87	37.53	4.94	4.62
Science	8	Winter	287,102	4.82	17.09	0.25	2.03	30.26	36.77	4.24	4.53
Science	8	Spring	301,412	4.79	16.48	0.23	1.89	29.91	36.47	5.32	4.93
Science	9	Fall	60,686	2.78	8.44	0.50	3.79	27.66	44.03	9.48	3.34
Science	9	Winter	47,129	2.29	9.16	0.55	4.02	29.14	42.66	9.30	2.88
Science	9	Spring	52,219	3.73	7.80	0.51	3.69	27.75	43.73	9.08	3.73
Science	10	Fall	56,559	3.60	9.05	0.40	3.86	29.45	41.14	9.23	3.28
Science	10	Winter	45,128	3.19	9.77	0.49	3.96	30.65	39.98	9.01	2.95
Science	10	Spring	45,846	3.34	8.47	0.49	3.71	27.79	42.09	10.12	4
Science	11	Fall	39,046	3.34	9.14	0.62	4.12	28.99	41.06	9.70	3.03
Science	11	Winter	30,351	2.74	9.92	0.59	4.29	30.85	39.89	8.98	2.72
Science	11	Spring	26,318	2.22	9.43	0.65	4.61	26.55	41.65	10.72	4.18
Science	12	Fall	15,643	4.56	7.35	0.79	5.25	36.38	31.78	10.04	3.84
Science	12	Winter	11,399	3.91	7.65	0.73	5.87	35.89	32.45	9.96	3.54
Science	12	Spring	9,088	3.75	7.98	0.73	5.01	35.07	30.72	10.95	5.80

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