map®
SPANISH

Spanish MAP Growth Reading Technical Report

March 17, 2021

nwea

Acknowledgements: It is with great appreciation that we recognize the many people at NWEA who contributed to this technical report. It was a collaborative effort involving people from numerous departments in the organization. We give special thanks to those who conducted the analyses and wrote and edited the document, including Carmen Hall, Teresa Krastel, Kelly Rivard, Jeff Luther, Patrick Meyer, Shudong Wang, Adam Withycombe, and countless others whose expertise and knowledge about Spanish MAP Growth was crucial.

Table of Contents

Executive Summary 7

Chapter 1: Introduction 10

1.1. Background 10

1.2. Pilot Overview 11

1.3. Intended Purpose and Uses of Test Scores 12

Chapter 2: Test Design 13

2.1. Content Design and Structure 13

2.2. Item Alignment to Standards 14

2.3. Test Construction 15

2.4. Test Content Validation 16

Chapter 3: Content Development 17

3.1. Item Types 17

3.2. Item Specifications 19

3.3. Cognitive Complexity 19

3.4. Transadaptation Process 19

3.5. Item Development 21

3.5.1. Item Writing 21

3.5.2. Item Review 21

3.6. Passage Development 22

3.7. Text Readability 23

Chapter 4: Test Administration and Security 25

4.1. Adaptive Testing 25

4.2. Test Engagement Functionality 26

4.3. User Roles and Responsibilities 26

4.4. Administration Training 27

4.5. Practice Tests 27

4.6. Accommodations and Accessibility Features 28

4.6.1. Universal Features 28

4.6.2. Designated Features 29

4.6.3. Accommodations 30

4.6.4. Third-Party Assistive Software 31

4.7. Test Security 31

4.7.1. Assessment Security 33

4.7.2. Role-Based Access 33

Chapter 5: Scale Development, Scoring, and Item Calibration 34

5.1. Rasch Unit (RIT) Scales 34

5.2. Scaling 35

5.3. Field Testing 36

5.4. Field Test Item Evaluation 38

5.4.1. Item Fit 39

5.4.2. Model of Man (MoM) Procedure 40

5.4.3. Human Review Process 40

Chapter 6: Reporting 42

6.1. MAP Growth Reports 42
6.2. Learning Continuum 44
Chapter 7: Reliability 45
7.1. Test-Retest Reliability 45
7.2. Marginal Reliability (Internal Consistency) 46
7.3. Score Precision 47
Chapter 8: Validity 49
8.1. Evidence Based on Test Content 49
8.2. Evidence Based on Internal Structure 49
8.2.1. Goodness-of-Fit Indices 49
8.2.2. Results 51
8.3. Differential Item Functioning (DIF) 53
Chapter 9: Pilot User Norms 56
9.1. Pilot User Norms vs. National Norms 56
9.2. Estimating the Pilot User Norms 57
9.3. Norms Reference Groups 57
9.4. Achievement Status and Growth Norms 58
References 60
Appendix A: Spanish Pilot User Norms 63

List of Tables

Table 1.1. Available Spanish MAP Growth Assessments 10
Table 2.1. Alignment Guidelines for Spanish MAP Growth 14
Table 3.1. Item Types included on Spanish MAP Growth Reading Assessments 17
Table 3.2. Quantitative and Qualitative Analyses 24
Table 4.1. User Roles in the MAP Growth System 26
Table 4.2. Available Universal Features 28
Table 4.3. Available Designated Features 29
Table 4.4. Available Accommodations 30
Table 4.5. Third-Party Assistive Software 31
Table 4.6. Test Security Before and During Testing 32
Table 5.1. Field Test Plan 36
Table 5.2. Number of Pilot Test Events 37
Table 5.3. Pilot Student Sample Demographics 38
Table 5.4. Fit Index Descriptions and Criteria 39
Table 6.1. Required Roles for Report Access 42
Table 6.2. Report Summary 42
Table 7.1. Test-Retest with Alternate Forms Reliability by Grade 46
Table 7.2. Marginal Reliability by Grade 47
Table 8.1. Goodness-of-Fit Summary Indices 51
Table 8.2. DIF Categories 54
Table 8.3. Number of Items in DIF Study and Descriptive Statistics of DIF Student Sample 55
Table 8.4. DIF Results 55

Table 9.1. Pilot Spanish User Norms vs. English National Norms 56
Table 9.2. Summary Descriptive Statistics of Sample Test Scores..57
Table 9.3. Status and Growth Norms for Spanish MAP Growth Reading Test of Grade 1.........59

List of Figures

Figure 3.1. Sample Item—Multiple-Choice 18
Figure 3.2. Sample Item—Choice Multiple 18
Figure 3.3. Sample Item—Selectable Text 18
Figure 3.4. Sample Item—Drag-and-Drop..18
Figure 5.1. Reading Item with Good Model Fit 40
Figure 7.1. SEM of RIT Scores 48

List of Abbreviations

Below is a list of abbreviations that appear in this technical report.

AOR	Aspects of Rigor
ASG	Achievement Status and Growth
CCSS	Common Core State Standards
CCSSO	Council of Chief State School Officers
DIF	differential item functioning
DOK	Depth of Knowledge
ELA	English Language Arts
ELL	English language learner
ETS	Educational Testing Service
HLM	hierarchal linear model
IEP	Individualized Education Program
IRT	item response theory
MH	Mantel-Haenszel
MLE	maximum likelihood estimation
MoM	Model of Man
MSE	mean square error
RIT	Rasch Unit
RMSE	root mean square error
SCI	School Challenge Index
SD	standard deviation
SEM	standard error of measurement
TTS	text-to-speech
UDL	Universal Design for Learning

Revision History

Date	Version	Description
12/3/2019	0.1	Initial draft created by Shudong Wang
6/24/2020	0.2	Major contributions by subject matter experts
3/17/2021	1.0	Finalized by Patrick Meyer; published

Executive Summary

This technical report documents the processes and procedures employed by NWEA® to build and support the Spanish MAP® Growth™ Reading assessment. It is written for measurement professionals and administrators to help evaluate the quality of the assessment. Principal information presented in each chapter is summarized below. This report is not intended to be an administration guide or a technical description of the hardware and software needed for use of the system. For additional information not covered in this technical report, please contact your local NWEA representative or consult the NWEA website at www.nwea.org.

Chapter 1: Introduction

This chapter summarizes the Spanish MAP Growth Reading assessment and provides the intended uses of test scores. The purpose of the Spanish MAP Growth Reading assessments is to help districts, schools, and teachers better understand what Spanish-speaking students know and are ready to learn. NWEA accumulated pilot test data to support development of the assessment, produce pilot user norms for these Spanish reading tests, and to create the Spanish reading scale. Pilot data were collected for K–5 students in fall, winter, and spring of 2018–2019. Pilot data for Grades 6–8 students were captured in Spring 2019 only. The general release in Fall 2019 was available to all existing and new partners. All participants volunteered to take the tests. Any partner that was interested could join the study, although the target students were native Spanish speakers receiving both English and Spanish instruction and who would take both English Reading and the Spanish Reading pilot tests in the same term. The pilot assessment included an adaptive component with field test items selected using a goal-balancing method. While field test items do not traditionally count toward students' scores, the adaptive field test items in the pilot were labeled as operational and included provisional Rasch Unit (RIT) scores for reporting and research purposes. The pilot tests featured both transadapted and newly developed native Spanish items.

Chapter 2: Test Design

This chapter summarizes the design of the Spanish MAP Growth Reading assessments that have a parallel structure to the English version to be able to make comparisons between the two tests, allowing for the creation of a Spanish MAP Growth Reading vertical scale and the ability to link both the Spanish and English versions of the assessment.

Chapter 3: Item Development

This chapter describes the item types and the item development and review processes for the Spanish Reading pilot. About 2,250 native Spanish items were developed across Grades K–8, and about 1,350 items were transadapted from the English pool. The goal was to then develop items that address the Spanish standards that transadaptation did not cover. For the transadapted items, original English content was adapted to be culturally and linguistically appropriate in Spanish. Items selected for transadaptation address the content but are either culturally neutral or have appropriate parallel terms and concepts that can be adapted to the Spanish language and culture for fairness and accuracy. Item writing and review for the Spanish MAP Growth Reading assessments followed a very similar process as the English items.

Chapter 4: Test Administration and Security

This chapter describes the test administration and test security processes. Similar to the English version of MAP Growth, the Spanish MAP Growth assessments are fully adaptive, and each student experiences a unique test based on their responses to each item. Spanish MAP Growth Reading 2–8 tests take about 46–60 minutes, and K–2 assessments take about 40 minutes. The assessments can be administered up to four times a year (fall, winter, and spring, with a fourth optional administration in summer). Access to the MAP Growth system is based on differentiated roles such as system administrator and proctor. Practice Spanish MAP Growth Reading tests are available that provide the same access and functionality as the actual tests. The assessments have several features to improve test fairness and provide more precise and valid measurement, including universal features such as a calculator and highlighter, designated features such as text-to-speech (TTS), and accommodations such as assistive technology. Test security for Spanish MAP Growth follows the same process as their English counterpart.

Chapter 5: Scale Development, Scoring, and Item Calibration

This chapter describes the development of the RIT scale, the scaling process, item calibration, and the evaluation of field test items. RIT scores range from 100 to 350 and are on an equal-interval, vertical scale than spans multiple grades. Spanish MAP Growth Reading is a parallel assessment with a link to the existing English MAP Growth Reading assessments. English and Spanish MAP Growth Reading assessments measure similar but not identical reading constructs, so NWEA statistically linked the scales of the English and Spanish versions of MAP Growth based on a bilingual group design to allow comparisons of student scores across assessments to create the Spanish vertical scale and link student scores across languages. Pilot field testing occurred in Fall 2018, Winter 2019, and Spring 2019. Good item parameter estimates are critical to the validity of a test based on IRT. Field test items are checked for model fit via item fit statistics, the Model of Man (MoM) procedure, and human reviews.

Chapter 6: Reporting

This chapter summarizes the score reports available at the student, class, and district levels. Report types include the Student Profile, Student Progress, Achievement Status and Growth (ASG), Class Breakdown by RIT, District Summary, and Skills Checklists and Screening reports. The learning continuum shows the content a student can encounter throughout the test by instructional area, standards, and RIT bands. This report can be used to show what students performing at a given RIT level have achieved and what they are typically ready to learn. It has two views: the class view and test view.

Chapter 7: Reliability

This chapter summarizes the reliability evidence provided for MAP Growth. Reliability refers to the consistency of achievement estimates obtained from the assessment. The reliability of the Spanish MAP Growth Reading assessments was examined via test-retest reliability, marginal reliability (internal consistency), and score precision based on the standard error of measurement (SEM). Data included in these analyses were from the Fall 2018, Winter 2019, and Spring 2019 administrations of the pilot data. Test-retest reliability coefficients range from 0.50 at Grade K to 0.83 at Grade 5. There is no test-retest reliability for Grades 6–8 because only one test administration was taken by these grades. The overall marginal reliabilities for all grades are in the 0.90s, which suggests that the tests have high internal consistency. Regarding score precision, the MAP Growth adaptive test algorithm selects the best items for each student, producing a significantly lower SEM than fixed-form tests.

Chapter 8: Validity

Validity is defined as the “the degree to which evidence and theory support the interpretations of test scores for proposed uses. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests” (AERA et al., 2014, p. 11). This chapter summarizes evidence based on test content and internal structure, including goodness-of-fit indices and differential item functioning (DIF). Overall, the goodness-of-fit results indicate that the constructs measured by the Spanish MAP Growth Reading tests across language background groups are at least tau-equivalent and most of them are parallel equivalent. The major implication of these results is that Spanish MAP Growth Reading tests can be used for students with different language backgrounds who receive different classroom instruction. Overall, DIF results show the following three patterns: (1) Most items are classified as A, (2) the highest percentage of C DIF is the Native English/Bilingual group (6.66%), and (3) C DIF is rare for the remaining DIF study groups (~1%).

Chapter 9: Pilot User Norms

This chapter summarizes the development of the pilot user norms. As of the July 2019 release of the Spanish MAP Growth Reading test, pilot user norms were made available for Spanish Reading within MAP Growth reports based on the pilot year of test data. Although they are drawn from a limited pool of test events and not a nationally representative like the general MAP Growth norms, they provide basic contextual information about student performance in the fall and spring and growth between fall and spring on the Spanish MAP Growth Reading assessments. NWEA intends to refresh the Spanish MAP Growth Reading user norms data for Fall 2020 based on available testing data pool from the 2019–2020 school year.

Chapter 1: Introduction

The English language learner (ELL) population is growing (4.6 million students in 2014–2015) and most ELL students speak Spanish as their native language (77.1% or 3.7 million), according to the National Center for Education Statistics (NCES; 2019). Bilingual and dual-immersion programs are also on the rise, especially in border states like California and Texas. Having trustworthy data about every student's performance allows educators to support equity in the classroom and better inform instruction. That means getting accurate scores for students who either speak Spanish as their first language or receive instruction in Spanish. It is essential for NWEA to meet the needs of Spanish-speaking students to support the mission of helping all students learn.

The purpose of the Spanish MAP Growth Reading assessments is to help districts, schools, and teachers better understand what Spanish-speaking students know and are ready to learn. To provide additional context and better support interpretation of student scores, NWEA accumulated pilot test data to support development of the assessment, produce pilot user norms for these Spanish reading tests, and to create the Spanish reading scale. Pilot data were collected for K–5 students in fall, winter, and spring of 2018–2019. Pilot data for Grades 6–8 students were captured for Spring 2019 only.

1.1. Background

As shown in Table 1.1, NWEA offers Spanish language versions of MAP Growth Mathematics and Reading tests that are parallel to the English versions.[1] They are adaptive tests that can be administered up to four times per calendar year in Grades K–2 and 2–5 and up to three times per calendar year in Grades 6+. While the Spanish MAP Growth Mathematics assessments use the same scale as the English mathematics assessments, Spanish MAP Growth Reading is on its own scale. Reporting features are also consistent with the English version. Spanish MAP Growth Reading is a parallel assessment with a scale linked to the existing English MAP Growth Reading assessments. Educators can receive scale score data from both English and Spanish MAP Growth Reading growth measures if students take both assessments, allowing them to make informed decisions to support their students' learning in both languages.

Table 1.1. Available Spanish MAP Growth Assessments

Test Type	Purpose	Testing Frequency	Content Area
Spanish MAP Growth Mathematics (Grades K–2 and 2–12)	Instructional areas are identical to the English MAP Growth Mathematics test. The Spanish version uses the same scale and reporting as the English version so that scores are consistent and comparable.	Up to four times per year in Grades K–2 and 2–5. Up to three times per year in Grades 6+.	• Mathematics
Spanish MAP Growth Reading (Grades K–2, 2–5, and 6-8)	Spanish version of MAP Growth Reading that measures achievement of students who either speak Spanish as their first language or receive instruction in Spanish. Piloted in 2018–2019 for operational use beginning in Fall 2019.	Up to four times per year in Grades K–2 and 2–5. Up to three times per year in Grades 6–8.	• Reading

[1] Details about the Spanish MAP Growth Mathematics are currently not provided in this technical report but will be in future iterations.

In Fall 2018, NWEA added Spanish MAP Growth Mathematics for K–2 to add to the already-established Spanish MAP Growth Mathematics for Grades 2–8. The Spanish MAP Growth Reading pilot began in Fall 2018 with the goal of gathering data to support the full release of an operational adaptive assessment in Fall 2019. The K–2 and 2–5 pilots were available in Fall 2018, and the 6–8 pilot began in Spring 2019. The general release in Fall 2019 was available to all existing and new partners.

1.2. Pilot Overview

A key purpose for conducting the pilot study was to create the Spanish MAP Growth Reading scale. Other purposes include field testing the Spanish MAP Growth Reading items to create an operational item pool and develop user norms. NWEA reached out to existing partners to participate in the Spring 2019 Spanish MAP Growth Reading pilot. All participants volunteered to take the tests, and no sampling design was involved. Any partner that was interested could join the study, although the target students were native Spanish speakers receiving both English and Spanish instruction and who would take both English Reading and the Spanish Reading pilot tests in the same term.

The pilot tests were available and aligned for the California Common Core State Standards (CaCCSS) en español, the Common Core State Standards (CCSS)[2] Spanish Language version (CCSS en español; National Governors Association Center for Best Practices & Council of Chief State School Officers [CCSSO], 2012), and Texas Essential Knowledge and Skills for English Language Arts and Reading. Aligning to these three standard sets provided a base to support a large number of Spanish-speaking students. California is also a CCSS state, but a few additional standards are California-specific, and the Spanish version of the standards covers Spanish-only content. Texas has state requirements for Spanish assessments.[3]

The 2019 pilot assessment included an adaptive component with field test items selected using a goal-balancing method. While field test items do not traditionally count toward students' scores, the adaptive field test items in the pilot were labeled as operational and included provisional Rasch Unit (RIT) scores for reporting and research purposes. The provisional RIT scores were based on their English counterpart items that were used as anchors to place the native Spanish items via professional judgment on level of difficulty. Pilot score reports were not counted as an official growth measure. They were generated for informational purposes only and reviewed for accuracy and appropriateness in preparation for the general release of the Spanish MAP Growth Reading test in Fall 2019. The assessments had a similar test length as the English version, were untimed, and were estimated to take approximately 45 minutes to complete. The pilot tests featured both transadapted and newly developed items. Specifically, the item pool included the following:

- Items that were transadapted from the MAP Growth Reading English item pool (i.e., translated and modified for cultural and linguistic appropriateness and checked for bias)
- Native Spanish items developed specifically for these tests

[2] © Copyright 2010 National Governors Association Center for Best Practices and Council of Chief State School Officers. All rights reserved.
[3] http://ritter.tea.state.tx.us/rules/tac/chapter128/ch128a.html

For transadapted items, standard alignments were the same as the counterpart English item. For newly developed items, items were written to evidence statements and specifications for the assessable standards in the MAP Growth Reading Spanish scale. California, Texas, and CCSS all have standards that address Spanish Literacy, some of which address skills specific to Spanish such as Reading Foundational Skills, Spelling, Conventions, and word parts.

1.3. Intended Purpose and Uses of Test Scores

The purpose of the Spanish MAP Growth Reading assessments is to measure bilingual and monolingual students' Spanish reading achievement and track longitudinal growth of reading achievement in Grades K–8. In general, MAP Growth assessment data can be used in numerous ways to support student growth and achievement. NWEA supports the use of MAP Growth scores to:

- Monitor student achievement and growth over time, from Grades K–8
- Plan instruction for individual students and groups of students at the classroom, grade, school, and district levels
- Compare student performances within normed groups
- Evaluate programs and conduct school improvement planning
- Summarize scores for district- or school-level resource allocation
- Combine RIT scores with other information (e.g., homework, classroom tests, state assessments) to make educational decisions
- Compare student performances between Spanish and English reading assessments to determine academic needs in each language

Chapter 2: Test Design

The Spanish MAP Growth Reading assessment was designed to have a parallel structure to the English version to be able to make comparisons between the two tests, allowing for the creation of a Spanish MAP Growth Reading vertical scale and the ability to link both the Spanish and English versions of the assessment. The sub-area structure and standard mappings match the English counterpart. The design of the Spanish MAP Growth assessments was guided by the same underlying principles as the English version, including the Universal Design for Learning (UDL) principles (Thompson et al., 2002).

When investigating the degree of construct equivalence between English literacy and Spanish literacy, the research literature points to significant differences with the learning-to-read space (Jiban, 2017). These largely derive from the extremely complex and inconsistent orthography in English, as compared with the shallow or transparent orthography of Spanish. Linguistic differences suggest that Spanish may bring shifts in relative difficulty across phonics, phonological awareness, spelling, and achievement of accurate decoding, as compared with English. By contrast, both vocabulary and reading comprehension may offer closer comparability, especially by the end of Grade 3. An important caveat is that achievement levels in these areas differ systematically for monolinguals as compared with emerging bilinguals. These broad findings are borne out by both research on reading and by reviews of English-specific and Spanish-specific standards.

2.1. Content Design and Structure

Spanish MAP Growth Reading assesses Spanish language and literacy. For Grades 2–8, tests on the Reading scale address reading comprehension, understanding of genres and text, and vocabulary. The Grades K–2 tests also cover Reading and some elements of Language Usage such as grammar, mechanics, and the elements of writing, as well as foundational skills (phonics, phonological awareness, and concepts of print) in addition to the comprehension and vocabulary consistent with 2–5. At this time, there is no Spanish MAP Growth assessment for 2–5 or 6+ on the Language Usage scale.

Each Spanish MAP Growth test is defined by the content area and grade band. Spanish MAP Growth Reading is broken into K–2, 2–5, and 6–8 tests. The K–2 test provides targeted Spanish audio support and addresses skills appropriate for students who are learning to read, including Reading Foundational Skills and Language and Writing standards. In contrast, students who take the 2–5 and 6–8 tests have progressed to independent reading. The split between the 2–5 and 6–8 test helps ensure that students see content appropriate to their age and achievement level. For example, when taking the 6–8 test, middle school students reading below grade level will see texts that allow them to demonstrate their reading skills without including overly juvenile references that may be perceived as demeaning. Similarly, advanced elementary readers will be challenged with increasingly complex texts without encountering excerpts from texts for which they have no frame of reference.

Within each test, the content is further defined by instructional areas that are derived from the structure of the content standards and provide information about how the content area is represented in the test. The instructional areas act as reporting categories. Each instructional area is further divided into sub-areas as another layer of defining the test content. The test samples evenly across all instructional areas, ensuring breadth of coverage of the standards.

2.2. Item Alignment to Standards

To perform alignment to the appropriate standards, NWEA content specialists crafted alignment guidelines tailored to the structure of the standards based on a review of supporting documents. An item was considered aligned when the item targeted either the whole standard or an integral part of a standard in a way that is both grade-appropriate and at a level of cognitive complexity addressed by the standard. Table 2.1 presents the Spanish MAP Growth alignment guidelines.

Table 2.1. Alignment Guidelines for Spanish MAP Growth

Approach to:	Spanish Reading
Definition of an aligned item	A student needs to demonstrate the knowledge and/or skill expressed* in the standard to respond correctly to the item. The student cannot or most likely cannot answer correctly without that knowledge and/or skill. The item may address the whole standard or a part of the standard in order to best focus on a single skill, a single portion of significant content, and/or a single cognitive level within the standard.
Assessable and non-assessable standards	NWEA only aligns to standards that have been defined as assessable. Assessable standards are often the most granular standards, although exceptions are noted below. Standards are only marked as assessable if they are appropriate for interim/formative assessment, NWEA has the functionality to assess them, and they are intended to be used on current blueprints. • Standards apply to the most granular that they can based on the nature of the standards. For example, for states that do not have a separate set of standards for Spanish language and literacy, alignment occurs at the superordinate standard. Therefore, in some cases alignment may not be the most granular. As those states have English-only standards at the most granular level, some Spanish items would not apply there. These are typically items that are aligned to the most granular standards of the CCSS Spanish standards). • Skills that are impractical for NWEA products (e.g., lengthy multi-part tasks that require longer than a normal class period) are not marked assessable. However, some standards (such as in writing, oral responses) are considered assessable via an approximation. • Parent standards are generally marked as non-assessable, although some parent standards are marked as assessable in Spanish K–2 for the reasons stated above. • The inclusion of audio in MAP Growth K–2 allows for assessment of standards in Reading: Foundations and some listening standards from the Speaking and Listening strand. • Standards requiring students to produce oral responses are assessed in a manner befitting a computer-adaptive assessment because these items still provide valuable information to teachers about students' knowledge of specific skills.
Prerequisite skills, related content, and implied content	• Items assessing prerequisite skills and/or content are not aligned. • Implied content is often open for interpretation. Therefore, content teams must make decisions and document those decisions for specific standards that are open to interpretation. Decisions must be based on deep consideration of the standard, standard set, and available resources from experts. • The term “e.g.” indicates examples of the type of content/skills that could fulfill the standard, but it is not an exhaustive list and the listed examples are not required to be assessed. The term “i.e.” indicates a rewording of the standard and therefore defines the limits of the content/skills that are included as an integral part of the standard. • If a standard says *including*, it means the content must be included when assessing that entire standard (it does not all have to be included in a single MAP Growth item, though); when *such as* is used, it has a similar meaning as e.g.
Cognitive verbs/ cognitive expectation in a standard	The cognitive verbs are closely considered as the primary indication of the cognitive expectation associated with a given standard. Items that do not meet that cognitive expectation should not be aligned. However, some standards, most notably writing, are assessed via an approximation that does not meet the expectation or exact action encompassed by the cognitive verb. Decisions should be clearly documented. This can be more difficult to achieve with non-CCSS standard sets.

Approach to:	Spanish Reading
Granularity of alignment (e.g. parent/child, anchors, clusters)	Align to most granular portion of standard except in cases noted below. • Spanish MAP Growth Reading 2–5 and 6–8 do not align items to parent standards, although Spanish K–2 items can align to parent standards (as noted above). • For ELA, NWEA recognizes the special assessability concerns around the standards CCSS designates as Language Progressive skills. NWEA has items targeting these progressive skills not only when they are first introduced but also at subsequent grades in accordance with the CCSS grade recommendation. Because CCSS has no codes or ways to directly note that alignment at the higher grades, NWEA uses the overarching/parent standards (L.1, L.2, and L.3) to align items assessing these progressive skills at higher grades. • Many CCSS-based standard sets do not adopt this aspect of the CCSS.
Alignment to the whole standard or portions of a standard	If possible, alignment would be to the entire standard. However, when standards are broad or complex, single items can target portions of a standard.
Grade-level considerations	• Items with distractors that have content that is above grade level should be aligned to a higher grade-level standard, if at all. • A holistic determination of grade level must be made that considers vocabulary, context, complexity of the task, readability of the text, and the content included in distractors. • The text in an item must be sufficiently complex for the grade level for it to fully align to that grade's standard. Consequently, for items in common stimulus passage sets, the text complexity of the passage is always considered. • The Reading passage asset adheres to quantitative (Spanish Lexile®) text complexity and qualitative (conceptual appropriateness) measures as appropriate for the grade/grade band indicated in the item specifications.
Basis for alignment decisions	Alignment decisions are based on information and resources obtained from the CCSS Translation Project website. This includes the appendices and other materials available at the sites. Additional resources provided by organizations closely involved with developing the CCSS en español, sample items from the consortia, and other vetted sources are also consulted.

*Content/skills should be directly stated or strongly implied. If implied, the acceptable content/skills should be documented by the content team, with decisions based on discussion and resources from expert sources.

2.3. Test Construction

Once NWEA content specialists have created instructional areas and sub-areas for a test, they align standard statements to these areas to establish the test structure and content. This combination of instructional areas, sub-areas, and standard statements is called a test blueprint. Once the blueprints are created, the MAP Growth item bank is reviewed, and appropriate items are aligned to the standards. These components form the eligible item pool for the test, along with the reporting structure and how all the eligible items fit into this structure. Additional constraints may be added to a test that may further limit the eligible item pool, including item selection requirements during test administration as required by the test type and item filters based on specific item metadata. These constraints are based on the target student population and may include item attributes such as item language or item accessibility for different student populations.

During test administration, the blueprint helps drive item selection to ensure that items presented to a student cover all instructional areas at a difficultly level appropriate to that student's performance, both overall and within each instructional area. Item selection is not restricted to items within a student's grade, allowing MAP Growth to better target students who are performing above or below the grade level mean for an instructional area. The test behavior during testing is defined in terms of the test length and item selection criteria for each section of the test as determined by the test content area and purpose. Once these elements are combined, the test is published to the testing platform as a defined set of behaviors and test metadata elements. Each item is also published to the testing platform, along with item metadata and information that determines to which tests the items belong. Tests go through a series of checks, including test content validation that simulate test runs of students at different ability levels, to ensure that the test item pools provide sufficient depth to cover the achievement continuum within each instructional area. Tests are then made available to specific partners based on their licensing agreements with NWEA.

2.4. Test Content Validation

Test content validation is performed as part of the broader process of aligning MAP Growth to different content standards and publishing new tests. The purpose of test validation is to ensure that each newly aligned MAP Growth item pool performs as intended. The process for the Spanish version of MAP Growth is similar to the process for the English version, although a main purpose of the pilot assessment was to create a valid Spanish MAP Growth Reading scale. Another difference is that the Spanish pilot validation used non-operational items in the pool because the Spanish scale was not created yet, whereas the English validation uses operational items in the pool. In general, test validation takes the form of test simulations with the operational item pool to determine the accuracy of student ability estimation and content coverage of an adaptive test. Tests are classified as pass, pass with qualifiers, or fail. Most tests pass or receive a qualified pass.

For the Spanish Reading pilot, an NWEA psychometrician conducted the test content validation simulation studies by following the steps below:

1. Simulate a MAP Growth adaptive test based on the operational item pool.
2. Simulate student growth over a two-year timeframe, typically six to eight administrations.
3. Apply longitudinal constraints that prevent a student from seeing the same item more than once in a set timeframe, typically 14 months (although the Spanish MAP Growth Reading longitudinal constraint is three months due to the current size of the item pool).

To determine if a test passes the validation, the psychometrician evaluates the following:

- Student ability estimation based on statistics including bias, mean square error (MSE), root mean square error (RMSE), and SEM. The better the estimation, the smaller these statistics will be.
- Content balancing based on how well the test meets the blueprints. A quality adaptive test should administer items among the instructional areas as stated in the blueprint.
- The efficiency of the adaptive algorithm based on the discrepancy between the interim ability estimate and item difficulty. The sooner the algorithm settles on the simulated student’s true ability value, the sooner the SEM criteria are satisfied.
- Item pool depth based on item RIT distribution at the overall test and instructional area levels. At each level, the pool should ideally span the full range of RIT values and have an adequate number of items at each RIT value to avoid running out of items.

Chapter 3: Content Development

About 2,250 native Spanish items were developed for the Spanish MAP Growth Reading assessment across Grades K–8, and about 1,350 items were transadapted from the English pool. These are items that apply to either English or Spanish language and literacy (e.g., identifying main idea). New native Spanish item development covered the scope of the standards but with initial developments targeting the Spanish-only standards such as foundational reading and vocabulary standards (e.g., accents) or standards that have a corresponding English standard but are addressed differently due to the differences in Spanish (e.g., sound-symbol correspondence, phonetics). The goal was to develop items that address the Spanish standards that transadaptation did not cover.

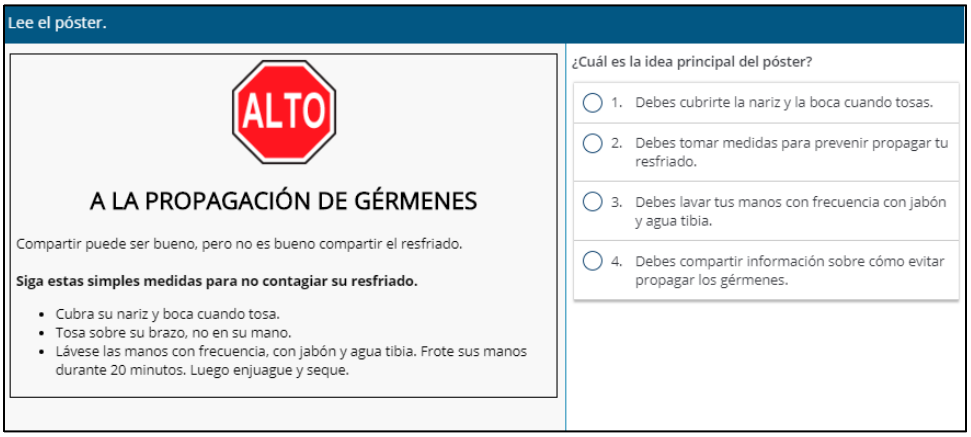
3.1. Item Types

Table 3.1 presents the item types included on the Spanish MAP Growth assessments. Figure 3.1 –Figure 3.4 present sample items.

Table 3.1. Item Types included on Spanish MAP Growth Reading Assessments

Item Type	Description
Multiple-Choice (Choice)	Students select one response from multiple options.
Multiple Select/Multiselect (Choice Multiple)	Students select two or more responses from multiple options. (Reading only)
Selectable Text (Hot Text)	Students select a response from within a piece of text or a table of information (e.g., word, section of a passage, number, symbol, or equation).
Drag-and-Drop	Students select an option or options in an area called the toolbar and move or “drag” these options (e.g., words, phrases, symbols, numbers, or graphic elements) to designated containers on the screen.
Click-and-Pop	Students move options (e.g., words, phrases, symbols, numbers, or graphic elements) from the area called the toolbar to designated container(s) on the screen by selecting an option; the option then “pops” into the container on screen.
Composite Items	Students interact with multiple interaction types included within a single item.

Figure 3.1. Sample Item—Multiple-Choice



Lee el póster.

ALTO

A LA PROPAGACIÓN DE GÉRMENES

Compartir puede ser bueno, pero no es bueno compartir el resfriado.

Siga estas simples medidas para no contagiar su resfriado.

- Cubra su nariz y boca cuando tosa.
- Tosa sobre su brazo, no en su mano.
- Lávese las manos con frecuencia, con jabón y agua tibia. Frote sus manos durante 20 minutos. Luego enjuague y seque.

¿Cuál es la idea principal del póster?

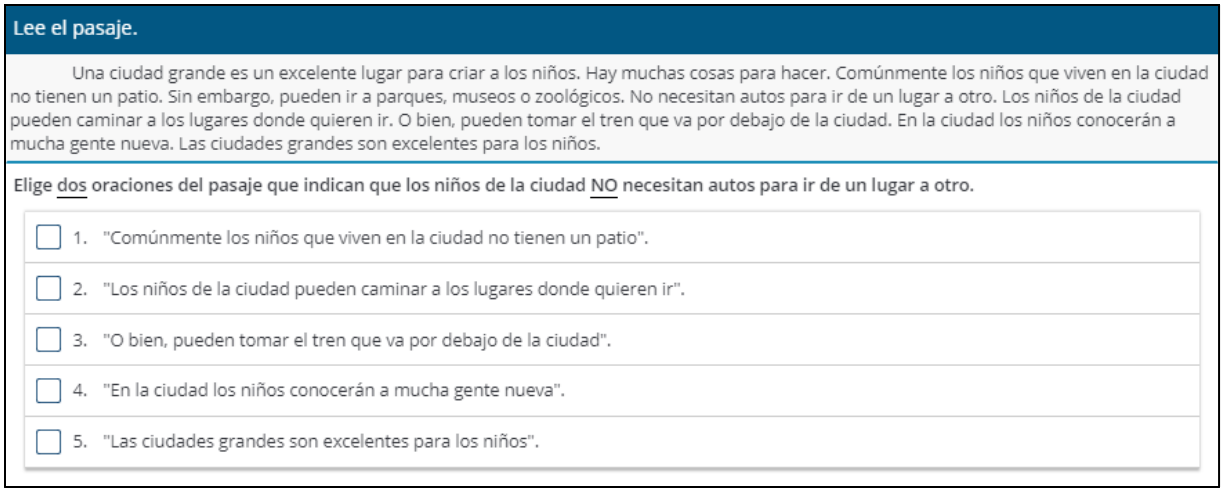
○ 1. Debes cubrirte la nariz y la boca cuando tosas.

○ 2. Debes tomar medidas para prevenir propagar tu resfriado.

○ 3. Debes lavar tus manos con frecuencia con jabón y agua tibia.

○ 4. Debes compartir información sobre cómo evitar propagar los gérmenes.

Figure 3.2. Sample Item—Choice Multiple



Lee el pasaje.

Una ciudad grande es un excelente lugar para criar a los niños. Hay muchas cosas para hacer. Comúnmente los niños que viven en la ciudad no tienen un patio. Sin embargo, pueden ir a parques, museos o zoológicos. No necesitan autos para ir de un lugar a otro. Los niños de la ciudad pueden caminar a los lugares donde quieren ir. O bien, pueden tomar el tren que va por debajo de la ciudad. En la ciudad los niños conocerán a mucha gente nueva. Las ciudades grandes son excelentes para los niños.

Elige dos oraciones del pasaje que indican que los niños de la ciudad NO necesitan autos para ir de un lugar a otro.

□ 1. "Comúnmente los niños que viven en la ciudad no tienen un patio".

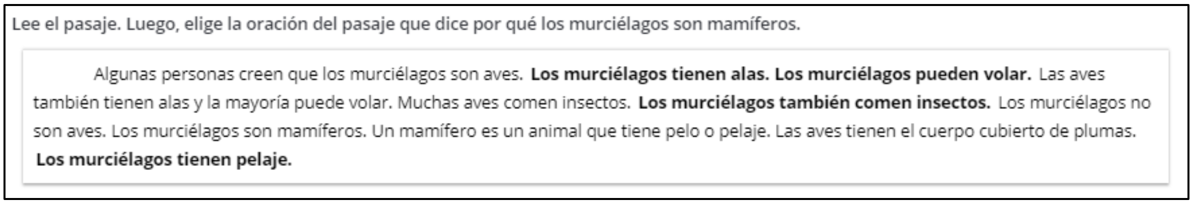
□ 2. "Los niños de la ciudad pueden caminar a los lugares donde quieren ir".

□ 3. "O bien, pueden tomar el tren que va por debajo de la ciudad".

□ 4. "En la ciudad los niños conocerán a mucha gente nueva".

□ 5. "Las ciudades grandes son excelentes para los niños".

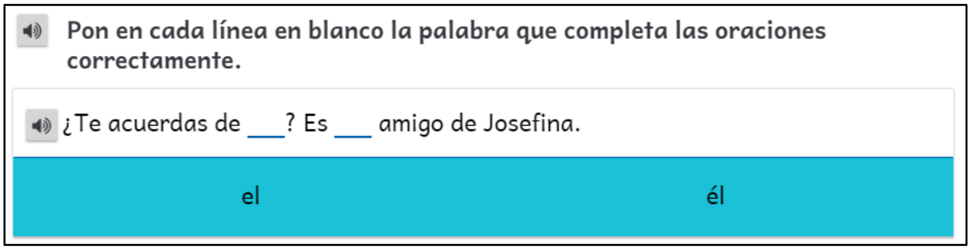
Figure 3.3. Sample Item—Selectable Text



Lee el pasaje. Luego, elige la oración del pasaje que dice por qué los murciélagos son mamíferos.

Algunas personas creen que los murciélagos son aves. **Los murciélagos tienen alas. Los murciélagos pueden volar.** Las aves también tienen alas y la mayoría puede volar. Muchas aves comen insectos. **Los murciélagos también comen insectos.** Los murciélagos no son aves. Los murciélagos son mamíferos. Un mamífero es un animal que tiene pelo o pelaje. Las aves tienen el cuerpo cubierto de plumas. **Los murciélagos tienen pelaje.**

Figure 3.4. Sample Item—Drag-and-Drop



Pon en cada línea en blanco la palabra que completa las oraciones correctamente.

¿Te acuerdas de ___? Es ___ amigo de Josefina.

el

él

3.2. Item Specifications

Item specifications are written to help content developers create items that are aligned to and assess an intended topic or skill. Spanish specifications are largely based on the ELA specifications but are Spanish-specific where applicable. Content specialists review each specification for clarity, completeness, and alignment to ensure that content developers will understand the types of items expected. NWEA item specifications, including both English and Spanish, are updated on an ongoing basis and include the following elements of guidance for item writers:

- Describe a direct and demonstrable relationship to areas of need
- Unpack an objective into discrete statements when the objective has numerous aspects
- Focus on one topic/skill and indicate a grade or grade range
- Ensure that no relevant skills are overlooked when unpacking an objective
- Match the cognitive complexity of the learning indicator
- Match the content to the item type based on best practices
- Provide guidance around passage/item resource/context when applicable
- Provide parameters, examples, definitions, and resources when applicable
- Provide suggestions on the types of answer choice options (e.g., the options for this item could be charts or graphs) when applicable

3.3. 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, Mathematics items have been tagged according to Student Achievement Partners’ Aspects of Rigor (AOR) model (Achieve, 2018).

3.4. Transadaptation Process

The Spanish MAP Growth assessments were developed to parallel the content and design of the English version. Many English items were selected and transadapted for these assessments. The Spanish items are not translations but rather transadaptations of the English items. Original English content was adapted to be culturally and linguistically appropriate in Spanish. Items selected for transadaptation address the content but are either culturally neutral or have appropriate parallel terms and concepts that can be adapted to the Spanish language and culture for fairness and accuracy. When identifying items for transadaptation, subject matter experts and content specialists considered the following:

- Language variation related to regionalisms and dialects
- Cultural difference within Spanish-speaking cultures and between Spanish- and English-speaking cultures
- Grammatical features and differences between the two languages
- Appropriateness of content (text and images) for the target Spanish audience and ability to preserve construct to target the skill of the original item for Spanish Language Arts

The International Test Commission guidelines informed the workflow and steps in transadapting the tests to help achieve similarity and parallelism in Reading (International Test Commission, 2017). The Reading tests assess similar concepts in either English or Spanish Language Arts in their respective languages. The transadaptation process aimed to

- define target audience and age group;
- define subject matter, test instrument, and topics;
- recruit linguists and subject matter experts for localization, regionalisms, and determination of need for more neutral, universal variants;
- confirm technical requirements;
- conduct localization review of sampling of items;
- compile glossaries, style guides, terminology; and
- achieve test equivalence.

To conduct the transadaptation, NWEA ELA and Spanish content specialists selected a possible pool of items to be translated and transadapted. The items did not have any copyright or public domain passages. A glossary of terms was created for consistency, and text extraction of both item and image text was used to translate and transadapt new terms and add to the glossary. Each item was analyzed to determine whether to translate, transadapt, or reject as untranslatable. After the translation and transadaptation of item content and images, items underwent three rounds of review: linguistic editing, content editing, and proofreading:

- Linguistic editing:
  - Edit for grammar, syntax, style, and flow
  - Ensure that concept and meaning were the same for both Spanish and source items (e.g., cognates, idiomatic expressions, adages)
  - Ensure that language and terms are appropriate and common for target audience
- Content editing:
  - Edit to correspond to target audience' Spanish educational background
  - Format to match source item
- Proofreading:
  - Use generic, neutral Spanish for Spanish speakers in the United States

To finalize the transadaptations, subject matter experts then validated the transadaptations and resolved issues. Translations underwent review and revision, if needed, before undergoing a final content review to ensure the quantity and quality of the content, preserve the knowledge and abilities needed to answer the item, ensure that item content represented curriculum of the target population, and ensure that the source item did not have errors, flaws, or multiple keys. Transadapted items were entered into the NWEA content management system, with images and appropriate audio and media requests. A final quality check was conducted to make sure the English and Spanish items looked the same and the metadata matched.

3.5. Item Development

Item writing and review for the Spanish MAP Growth Reading assessments followed a very similar process as the English items. Process summaries are provided below. For full details, please refer to the MAP Growth technical report.

3.5.1. Item Writing

NWEA content specialists develop items internally or contract out to vendors and 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, the developers are provided 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 item development resources such as item specifications and cognitive expectation frameworks that provide guidance regarding the content, context, cognitive complexity, and form of items. Content developers are also directed to an external documentation site with access to documents that provide guidance and requirements for the following:

- Item formatting and style
- Item type guidelines for when and how to construct a certain type of item
- Content-area-specific item writing guidelines
- UDL guidelines, including those for bias, sensitivity, fairness, and accessibility
- How to request media for items
- Copyright and permissions guidelines

NWEA selects freelance content developers and external vendors by following a strict vetting process that requires candidates to demonstrate expertise in their content area. NWEA requires that prospective content developers submit sample items in support of evidence in their resumes that they have the relevant content area knowledge, classroom teaching experience, and/or professional assessment writing experience. When there is a need for higher volumes of items, NWEA contracts with established content development vendors whose item samples are rigorously evaluated by NWEA content specialists and copyright and permissions specialists.

3.5.2. Item Review

Each item in the MAP Growth item pool undergoes the review process summarized below. While this process reflects the internal process conducted by NWEA, items also go through similar stages when developed by an external vendor. The difference lies in who owns the review stage, although each reviewer—whether at NWEA or from an external vendor—are highly qualified and have the appropriate content knowledge to perform the review. NWEA assists the external vendors as needed throughout the review process and always conducts final reviews prior to publishing.

In general, a minimum of three separate professionals (i.e., two content specialists and a copy edit/quality control specialist) thoroughly review each item. All items undergo a copyright and permissions review. An item can be sent back to a previous stage or rejected if it does not meet the strict standards of NWEA at any point during these reviews. All passages in the Spanish Reading bank are commissioned passages or transadaptations of commissioned passages from the Reading banks.

1. A copyright and permissions specialist ensures that public domain content is from authoritative, authentic sources and that content is free of plagiarism.
2. Content specialists ensure that the content is valid and meets the NWEA quality content and alignment standards. Content specialists also validate factual material, ensure that current topics are used, review for bias and sensitivity, and ensure instructional relevance. They also validate the grade appropriateness of the item and assign a DOK level and Bloom's classification.
3. A content specialist assigns a preliminary difficulty level (i.e., a provisional RIT) to newly developed items for field test purposes. Transadapted items were assigned the RIT of the English counterpart item.
4. The media developers create any graphics or audio required for an item. A copyright and permissions specialist ensures that the images and graphics do not resemble or infringe on trademarked, branded, or copyrighted images.
5. A copy editor reviews items for grammar, usage, and mechanics errors and ensures that the items adhere to style guidelines. The item is reviewed for visual bias, and image descriptions ("alt text") are added to graphics for use by screen readers. Image descriptions may allow students who use refreshable braille and/or screen readers to answer items that would otherwise be inaccessible. They also ensure that items display correctly in all supported browsers.

3.6. Passage Development

Text excerpts are used with Spanish MAP Growth Reading items. Some are short passages attached to standalone items, whereas others are extended texts that can support multiple items (i.e., common stimulus passages). To assess students' ability to analyze reading passages in a way that fully integrates the depth and breadth of academic reading standards, students need to engage in close reading of high-quality complex text of various genres and types. Common stimulus passages are presented with a set of several text-based items that require close reading of an extended text and are therefore included in the item bank to address concepts and state standards that require complex texts. The Spanish MAP Growth Reading item bank includes approximately 39 common stimulus passages that were either transadapted or commissioned. The K–2 assessment includes very short assets in standalone items and does not have common stimulus passages.

All Spanish passages were either transadapted commissioned passages or newly developed commissioned passages that include both literary and informational texts. Writers provided source documentation for informational texts, and all passages underwent a permissions and copyright review. The passage acquisition and review process for Spanish MAP Growth is similar to the process for the English version:

1. Content specialists write passage specifications to garner literary, informational, and persuasive passages, as well as technical, domain-specific, and historical documents. Specifications detail the desired readability, text complexity, word count, and genre.
2. External content developers fulfill passage specifications when submitting commissioned works.
3. Content developers send a synopsis of the passage topic to NWEA for preapproval. Before preapproving a topic, content specialists ensure that the topic is age- and grade-appropriate, does not overlap with topics of other passages, and is unlikely to present bias, sensitivity, or fairness concerns. Passage writers/finders submit passage files and relevant source documentation to NWEA.

4. All passages undergo a series of reviews conducted by copyright and permissions specialists, content specialists, and content production specialists. Reviews include the following tasks:
    a. Copyright and permissions specialist verifies that the passage is free of plagiarism (if commissioned).
    b. Copyright and permissions specialist ensures that the passage does not have copyright, trademark, or rights of publicity issues.
    c. Content specialist ensures that the passage meets the specifications and quality requirements and verifies that it meets the text complexity requirements for the grade level and is free of bias, sensitivity, and fairness issues. The content specialist also fact-checks commissioned informational passages.
    d. Content specialist reviews and revises commissioned passages to ensure accuracy and overall structural and mechanical quality and applies readability analysis to help gauge grade-appropriateness and quantitative text complexity.
    e. All passages are reviewed for bias, sensitivity, and fairness according to internal NWEA bias, sensitivity, and fairness guidelines. Content production specialists perform a final copyedit of commissioned passages to ensure that the passages conform to both NWEA-specific and publishing industry styles.

When evaluating texts, content specialists apply the following criteria:

- Expert and credible authorship: Does the author write with authority about the topic? What are the author's journalistic and academic credentials? Does the author have an authentic connection to the culture depicted in the work?
- Text worthy of study: Is the work well crafted? Does it lend itself to close reading and analysis? Does it contain a clear central idea, relevant evidence, opportunities for reasoning, concrete details, an effective structure, and rich and varied language?
- Text not widely taught: Is the text one that students are unlikely to have encountered in the classroom?
- Free of bias and sensitivity concerns: Does the text present people fairly, respectfully, and without stereotype?
- Inclusivity: Do the texts include diverse groups of people with diverse backgrounds and experiences?
- Engaging and appropriate for target readers: Is the topic and tone of the writing likely to appeal to students?
- Ideal for assessment: Does the text yield a variety of challenging, standards-aligned items?

3.7. Text Readability

The expected readability of text in items is specific to the item scale. NWEA content specialists evaluate the readability of passages using both quantitative and qualitative measures. Passages within a grade level are assigned a range of complexity: minimally complex, moderately complex, and highly complex. Table 3.2 presents the quantitative and qualitative analyses conducted for passages.

Table 3.2. Quantitative and Qualitative Analyses

Quantitative Analysis	• Research-based recommendations highlight the use of two or more quantitative text analyzers/readability measures. • NWEA captures several quantitative readability scores (e.g., Spanish Lexile®) for each passage. • While variation exists among text analyzers, no single measure is interpreted to outperform the others.
Qualitative Analysis	• Qualitative dimensions of a work are evaluated for developmental appropriateness, cognitive difficulty, and intended audience. • NWEA has developed an internal rubric used to evaluate passages on such criteria as Levels of Meaning, Structure, Language Convention and Clarity, and Knowledge Demand. • Qualitative analysis includes how information and ideas are communicated implicitly, such as through literary techniques like allusion or analogy. Also evaluated are reader's purpose, type of reading (surface level or deep analysis), and intended outcome (knowledge, solution, engagement, assessment).

Chapter 4: Test Administration and Security

Similar to the English version of MAP Growth, the Spanish MAP Growth assessments are fully adaptive, and each student experiences a unique test based on their responses to each item. Spanish MAP Growth Reading 2–8 tests take about 46–60 minutes, and K–2 assessments take about 40 minutes. The assessments can be administered up to four times a year (fall, winter, and spring, with a fourth optional 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 a lockdown browser. 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 things like pausing and resuming their test if needed.

The NWEA test delivery platform supports more than 60 million student test events each year. The platform has delivered uninterrupted service with 172,000 students actively testing, defined as “concurrent” users. The most recent configuration has been certified and tested for at least 300,000 concurrent users.

4.1. Adaptive Testing

The MAP Growth adaptive testing algorithm starts item selection using items with RITs that are as suitable as possible for a student’s abilities based on known information about the student (e.g., grade level, prior RIT scores). If the student answers the item correctly, they receive a more difficult item. An incorrect response prompts an easier item. Maximum Fisher’s information method is used for item selection coupled with a randomesque exposure control procedure that selects one out of a few items that can provide the most information about the student (Kingsbury & Zara, 1989).

To ensure test content validity and the comparability of different tests, a content-balancing procedure proposed by Kingsbury and Zara (1991) and commonly used in most adaptive tests is used. This content-balancing algorithm selects items from the most underrepresented content area according to its target administration value specified in the test blueprint. That is, once an item is administered by maximum information at the student’s current ability estimate, its content classification is evaluated against target values defined in advance in the test blueprint for each student. If the selected item represents a content area that is the least represented at that stage, this item is administered. The maximum likelihood estimation (MLE) method is used for final ability estimation.

Test length varies for different content areas. Tests terminate either when the maximum test length is reached or when final RIT scores meet the pre-specified measurement precision level. Struggling students who might otherwise get frustrated and stop trying and high-achieving students who might get bored by strictly grade-level assessments will remain interested as subsequent items adapt to their abilities.

4.2. 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. In July 2018, NWEA added a rule that invalidates tests when students show disengaged responses on 30% or more of items. However, in October 2018, NWEA rolled back the rapid-guessing invalidation rule but continued to alert proctors of rapid-guessing and provide rapid-guessing information. 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.
- To better support retesting processes, educators, including proctors, have access to reports showing students with invalidated tests due to excessive rapid guessing.

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 of difficulty to prevent this downward drift. After the student has rapid-guessed three items in a row, the proctor is notified 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.

4.3. User Roles and Responsibilities

Access to the MAP Growth system is based on multiple defined roles, as described in Table 4.1. Each role in the system 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 more roles can be assigned. For added security, the system requires manual steps to set up user accounts and authorization levels. Only users with data administrator or proctor permissions can create or modify student profiles. This limits the ability to change student information (e.g., demographics and class assignments) to authorized users who support roster preparation or test proctoring.

Table 4.1. User Roles in the MAP Growth System

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.

Role	Permissions & Responsibilities
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, this is likely a special education teacher or similar role. • View students within their school and add them to custom groups for instruction and reporting.

4.4. Administration Training

While there is no professional learning specific for Spanish MAP Growth, administration training is provided as part of the professional learning services provided by NWEA that includes in-person and online training professional development sessions. The process begins with a consulting session with an NWEA Professional Learning Consultant. NWEA then recommends four days of onsite professional learning, beginning with MAP® Growth™ Administration, Applying Reports, and MAP® Skills™ Basics workshops. During these sessions, educators learn to use MAP Growth; access, interpret, and apply MAP Growth data; and use the data to inform ongoing work, including goal-setting with students. An online MAP Growth administration workshop is also available that involves two three-hour sessions with 40 participants each who learn about administering the tests, accessing reports, and applying data.

4.5. Practice Tests

Practice Spanish MAP Growth Reading 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 that determines which practice tests the user will have access to. The username and password are both “grow.” Practice tests specifics are as follows:

- Not adaptive
- No score
- No proctor control

- Available in any supported browser and any supported device
- Available for multiple grades and content areas
- About five items depending on the grade

4.6. Accommodations and Accessibility Features

Spanish MAP Growth has several features to improve test fairness and provide more precise and valid assessment measurement. These features fall within three categories:

- Universal features
- Designated features
- Accommodations

Local schools and districts may determine whether certain features are considered universal, designated, or an accommodation. 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 both the student and educator.

4.6.1. Universal Features

Table 4.2 presents the available universal features for Spanish 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 4.2. 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 devices (such as a talking calculator or a braille calculator).
Highlighter	A student can mark desired text, items, or response options with a color.
Zoom	A student can increase the size of text and pictures onscreen.
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.
Keyboard navigation	A student can navigate through test content by using the keyboard (e.g., the arrow keys). This feature may differ depending on the testing platform.

Feature	Description
Non-Embedded	
Breaks (frequent breaks)	A student can take breaks, when needed, to reduce cognitive fatigue.
Dictionary	A student can use an English or Spanish dictionary, if necessary.
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	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.
Thesaurus	A student can use a thesaurus containing synonyms of terms.

4.6.2. Designated Features

Table 4.3 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 or teams using a consistent process. Embedded designated features are provided digitally through instructional or assessment technology. Non-embedded designated features (such as a magnification device) are provided locally.

Table 4.3. Available Designated Features

Feature	Description
Embedded	
Text-to-speech (TTS) (audio support, spoken audio)	A student can hear 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	A qualified human reader can read the test and item content out loud.
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.

Feature	Description
Native language translation	A test administrator who is fluent in the student’s native language can translate test and question 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.

4.6.3. Accommodations

Table 4.4 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 4.4. Available Accommodations

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

4.6.4. Third-Party Assistive Software

Third-party software features such as those in Table 4.5 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 4.5. 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. Windows 7 and 8 users can choose from either full screen or lens magnification modes.
Zoom on Mac and iPad	Mac computers and iPads have a built-in screen magnifier that can magnify a screen up to 40 times its normal display size.
Chromebook color contrast	High contrast mode inverts the picture so that a white background appears black, black text appears white, and colors are inverted (for example, blue text or graphics become orange).
Windows color contrast	Windows supports high contrast themes for the OS and apps that users may choose to enable. High contrast themes use a small palette of contrasting colors that makes the interface easier to see.
Mac and iPad color contrast	Increase the readability of the screen on your MacBook or iPad by increasing the contrast of the display. Increase the contrast of the whole screen or emphasize borders between items in the Display section of the Accessibility settings.
JAWS *Not yet available for Spanish	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.

4.7. Test Security

Test security for Spanish MAP Growth follows the same process as their English counterpart. 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 and Spanish MAP Growth that draw from large item pools offer several advantages for ensuring test and item security. The MAP Growth systems leverage the following inherent security advantages:

- 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.
- Longitudinal constraints are placed on assessments accordingly based on the size of the item pool to limit the number of times students see items. Spanish MAP Growth Reading assessments have a three-month longitudinal constraint (i.e., once a student has viewed an item, they will not see that item again for at least three months).
- Large item pools allow minor security breaches to be addressed by removing exposed items from the pool.
- Students within a program can easily be retested using a new set of items if there are questions about the integrity of their scores.

Other test security guidelines followed by NWEA include the following:

- 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.
- 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.
- 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.

The processes and tools provided in Table 4.6 are also used to ensure the integrity of the tests were not jeopardized, providing educators and students a positive and reliable user experience.

Table 4.6. Test Security Before and During Testing

Before test administration	• Rostering of student and educator data 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 with 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 possible with 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., ear phone for TTS). • Student test-taking is only allowed during the testing window. All tests are closed and access removed upon the close of testing window.

4.7.1. Assessment 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.

4.7.2. 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 to users needed to perform their jobs. 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.

Chapter 5: Scale Development, Scoring, and Item Calibration

MAP Growth items, including Spanish MAP Growth, are administered sequentially, with each item being selected to yield maximum information about student's ability. Individual tests are constructed based on the student's performance while responding to items constrained to a set of content standards. All items are dichotomously scored. RIT scores range from 100 to 350 and are on an equal-interval, vertical scale than spans multiple grades. Spanish MAP Growth Reading is a parallel assessment with a link to the existing English MAP Growth Reading assessments. Using the RIT scale to report test results makes it possible to follow a student's proficiency status across time. Changes in a student's score across administrations and years are interpreted as growth.

5.1. Rasch Unit (RIT) Scales

Development of the RIT scale was guided by item response theory (IRT) that rests on the relationship between student achievement and item characteristics (Lord & Novick, 1968; Lord, 1980; Rasch, 1960/1980). A benefit of using an IRT model is that student scores and item difficulties are on the same scale. The scale is equal interval in the sense that the difference between any two student scores is the same regardless of item difficulty. The same is true for the difference between any two item difficulties. The difference is constant throughout the scale.

Specifically, MAP Growth assessments use the one-parameter Rasch IRT model that estimates the probability (P_{ij}) that a student (*j*) with an achievement score of θ_j will correctly answer a test item (*i*) of difficulty δ_i. It is expressed as:

$$P_{ij} = \frac{e^{(\theta_j - \delta_i)}}{1 + e^{(\theta_j - \delta_i)}}. \tag{5.1}$$

The values of the achievement score and item difficulty in Model 5.1 are on the logit metric, an arbitrary scale commonly used for academic studies of the Rasch model. To allow the MAP Growth measurement scale to be easily used in educational settings, the following linear transformation of the logit scale is performed to place it onto the RIT scale developed by NWEA for use in all MAP Growth tests:

$$RIT = (\theta_j \times 10) + 200. \tag{5.2}$$

The RIT scale ranges from 100 to 350 and is not easily mistaken for other common educational measurement scales. The RIT scale, like other IRT measurement scales, has several useful properties when applied and maintained properly. The most important properties for the development of the measurement scales and item banks include the following, which have been empirically verified for the RIT scales (Ingebo, 1997) and can be used in a variety of test development and delivery applications:

- Item difficulty calibration is sample free (i.e., if different sets of students who have had an opportunity to learn the material answer the same set of items, the resulting difficulty estimates for an item are estimates of the same parameter that differ only in the precision of the estimate's value). The accuracy will differ due to the sample size and the relative achievement of the students compared to the difficulty of the items.

- Trait score estimation is sample free (i.e., if different sets of items are given to a student who had an opportunity to learn the material, the scores are estimates of the same student trait level). Again, precision may differ due to the number of items administered and the relative difficulty of the items compared to the student's level of achievement.
- The item difficulty values define the test characteristics. This means that once the difficulty estimates for the items to be used in a test are known, the precision and the measurement range of the test are determined.

Since IRT enables the administration of different items to different students while allowing for comparable results, the development of targeted tests becomes practical. Targeted testing is the cornerstone for adaptive testing. These IRT characteristics also facilitate the building of item banks with content that extends beyond a single grade or district, enabling the development of vertical scales such as the RIT scales that extend from kindergarten to high school.

5.2. Scaling

English and Spanish MAP Growth Reading assessments measure similar but not identical reading constructs. The difference in constructs exists in terms of linguistic differences across languages and achievement differences across developing stages. These differences affect scale development for the Spanish MAP Growth Reading assessment in two ways:

1. Not all items used in the Spanish Reading test are transadapted from the English version. Therefore, reading items that were not transadapted must be calibrated.
2. The K–2 assessment measures foundational skills, whereas the assessments for higher grades measure comprehension. Despite these differences in content, the K–2 assessment and the assessments for higher grades are considered to measures the Reading construct, which allows the development of a RIT scale that spans all grades.

The differences across grades in reading test content do not affect comparisons of student reading achievement between the Spanish and English versions because the two language versions of the assessment measure comparable content. Moreover, NWEA statistically linked the English and Spanish scales to allow for score comparisons between language versions of the assessment.

A bilingual group design (Sireci, 1997) was used to create the Spanish MAP Growth Reading scale from Grades K–8. It is a common person design where bilingual students took both the English and Spanish MAP Growth Reading tests by grade within approximately two weeks. The tests were randomly ordered by district (i.e., students took different versions of the test randomly across district). This design can achieve two purposes of scaling: horizontal linking across the English and Spanish MAP Growth Reading assessments and vertical linking across grades for Spanish MAP Growth Reading. It has an additional advantage of eliminating bilingual group differences in proficiency. Potential issues of this design are that bilingual students are not homogeneous regarding their native and second language proficiency and representativeness of the bilingual samples to either group of its monolingual cohorts may be skewed.

5.3. Field Testing

The item calibration method used for the Spanish Reading pilot study used the common person design. Steps for this design include (1) using English responses to score the student's English MAP Growth test and (2) fixing the student's English ability to calibrate items in the Spanish test. The field test data were used to address the following purposes:

- Create the Spanish MAP Growth Reading vertical scale
- Link the Spanish and English MAP Growth Reading scales
- Establish the Spanish Reading item pools
- Develop user norms descriptive of Spanish reading achievement in schools for given collected data

Pilot field testing occurred in Fall 2018, Winter 2019, and Spring 2019. For Grades K–5, the same students could take different tests across all three terms, whereas students in Grades 6–8 could only take the Spring 2019 test. The target samples were students who only spoke Spanish and students who could speak both Spanish and English. Students either participated in a monolingual or bilingual test administration, as shown in Table 5.1.

Table 5.1. Field Test Plan

		Administration Type	
Grades	2018–2019 Term	Monolingual	Bilingual
K–5	Fall, Winter, Spring	Spanish	Spanish + English
6–8	Spring	Spanish	Spanish + English

While only the bilingual data were used for scaling and linking, monolingual assessments were also offered because some students lack a minimum level of language proficiency to meaningfully participate in the English assessment. It therefore does not make sense to test these students in both languages. The monolingual test results were used to verify bilingual data. For the monolingual administration, students only took the Spanish MAP Growth Reading test. For the bilingual administration, students completed both the Spanish and English versions. The order of test administrations across both languages was balanced by allowing test administrators to decide the test order. This resulted in approximately half of students taking the Spanish test first and the other half taking the English test first. The overall effect of balancing test administration order is that student test carryover effects (e.g., memorizing items from one test to the next) can be reduced to a minimum.

Table 5.2 presents the number of student test events by grade, term, and language administration collected for this pilot study. NWEA collected 91,666 valid Spanish test events (based on the validation rules) across both administrations, with about half of the Spanish test events (53%) coming from the bilingual administration. All scaling, equating, and calibration samples were from bilingual test events to create the Spanish MAP Growth Reading user norms, whereas the Spanish scoring results were from the total number of Spanish test events.

Table 5.2. Number of Pilot Test Events

Grade	Term	#Spanish Test Events				
			Monolingual (Spanish Only)		Bilingual (Spanish & English)	
		Total	N	%	N	%
K	Fall 2018	4,140	2,942	71.1	1,198	28.9
	Winter 2019	4,763	2,870	60.3	1,893	39.7
	Spring 2019	5,862	3,064	52.3	2,798	47.7
1	Fall 2018	4,933	3,608	73.1	1,325	26.9
	Winter 2019	5,762	3,503	60.8	2,259	39.2
	Spring 2019	7,171	3,705	51.7	3,466	48.3
2	Fall 2018	5,217	3,113	59.7	2,104	40.3
	Winter 2019	5,318	2,684	50.5	2,634	49.5
	Spring 2019	7,038	3,290	46.7	3,748	53.3
3	Fall 2018	3,918	1,666	42.5	2,252	57.5
	Winter 2019	4,180	1,613	38.6	2,567	61.4
	Spring 2019	5,609	2,029	36.2	3,580	63.8
4	Fall 2018	3,318	1,289	38.8	2,029	61.2
	Winter 2019	3,722	1,447	38.9	2,275	61.1
	Spring 2019	4,947	1,671	33.8	3,276	66.2
5	Fall 2018	2,709	957	35.3	1,752	64.7
	Winter 2019	2,721	1,005	36.9	1,716	63.1
	Spring 2019	4,340	1,530	35.3	2,810	64.7
6	Spring 2019	2,371	313	13.2	2,058	86.8
7	Spring 2019	1,999	209	10.5	1,790	89.5
8	Spring 2019	1,628	252	15.5	1,376	84.5
	Total	**91,666**	**42,760**	**46.6**	**48,906**	**53.4**

Table 5.3 presents demographic information of students from the bilingual administration and from both the monolingual and bilingual administrations. In theory, desired proportional representation by groups across test administrations for generalization of test results should be as close as possible. As shown in Table 5.3, the overall distributions of student demographic information across the two types of test administrations (monolingual vs. bilingual) are very similar, indicating that the scaling, linking, and calibration samples adequately capture the demographic characteristics of the overall Spanish-speaking samples collected by NWEA in the pilot administration of the Spanish MAP Growth Reading assessment.

Table 5.3. Pilot Student Sample Demographics

Grade	N-Count	Gender*			Race and Ethnicity*								
		Female	Male	N/A	AI/AN	Asian	Black	Hispanic	NH/PI	White	Multiethnic	NS/Other	N/A
Bilingual Data													
K	2,345	49.30	50.70	0.00	3.71	0.17	2.35	51.64	0.04	40.60	1.45	0.04	0.00
1	2,725	50.31	49.69	0.00	2.97	0.29	1.80	61.47	0.04	31.30	1.80	0.33	0.00
2	3,343	49.84	50.16	0.00	3.44	0.12	1.62	60.36	0.09	29.70	2.99	1.68	0.00
3	2,920	50.07	49.93	0.00	2.77	0.24	0.72	64.21	0.17	29.32	0.38	2.19	0.00
4	2,688	49.37	50.60	0.04	3.68	0.11	0.48	64.73	0.00	29.09	0.41	1.49	0.00
5	2,180	49.59	50.41	0.00	2.71	0.18	1.28	64.22	0.28	30.37	0.50	0.46	0.00
6	874	47.14	52.63	0.23	2.40	0.11	0.80	91.99	0.00	4.12	0.23	0.34	0.00
7	645	48.68	51.32	0.00	0.93	0.16	0.47	93.64	0.00	3.88	0.16	0.78	0.00
8	578	43.08	55.71	1.21	1.90	0.17	0.35	93.60	0.00	3.29	0.17	0.52	0.00
Monolingual + Bilingual Data													
K	6,577	48.52	51.47	0.02	1.82	0.30	2.42	71.14	0.03	20.75	1.08	2.55	0.00
1	8,212	49.94	50.02	0.04	2.64	0.37	1.42	67.89	0.07	19.96	1.12	6.78	0.01
2	8,659	50.24	49.74	0.02	2.59	0.31	1.64	63.21	0.10	18.50	1.04	12.61	0.00
3	6,991	50.04	49.94	0.03	2.50	0.36	1.00	60.88	0.09	20.61	0.51	14.03	0.01
4	6,433	49.46	50.49	0.05	1.80	0.33	0.84	64.48	0.00	18.30	0.65	13.59	0.02
5	5,677	50.18	49.76	0.05	1.44	0.25	1.16	60.68	0.14	23.53	0.85	11.93	0.02
6	2,288	50.17	49.52	0.31	1.27	0.22	0.83	72.64	0.04	6.38	5.46	13.11	0.04
7	1,999	51.48	48.42	0.10	0.40	0.40	0.25	67.63	0.00	15.51	4.65	11.16	0.00
8	1,628	48.65	50.86	0.49	0.74	0.12	0.31	70.09	0.00	7.49	1.17	20.09	0.00

*N/A = Gender information is not available. AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. NS/Other = Not Specified or Other. N/A = Race and ethnicity information is not available.

5.4. Field Test Item Evaluation

Good item parameter estimates are critical to the validity of a test based on IRT. The evaluation of calibrated field test items ensures that the operational items work well with students. It also allows an opportunity for items to be reworded and field tested again to improve both the content and measurement quality of the item prior to being used operationally.

To evaluate a field test item's calibration, NWEA employs various descriptive statistics (e.g., percent correct, point-measurement correlation) and calculates item infit and outfit statistics that provide useful information about how well the responses adhere to the expectation of the Rasch model. However, various forms of information collected about an item's calibration status do not necessarily result in a decision about item quality. For example, some indicators can suggest good quality while others suggest caution. In such cases, human reviewers drive the final decision. However, human reviews are expensive and inefficient, especially when large numbers of items are under consideration. Recognizing this, NWEA adopts an integrated procedure called Model of Man (MoM) by employing automated procedures and human judgment. The automated procedure uses item fit statistics to mimic human review behavior and improve the overall quality and efficiency of the calibration process.

5.4.1. Item Fit

Item fit is evaluated with multiple indices and criteria, as shown in Table 5.4. Most of the indices provide information about the fit of the Rasch model to the observed responses. Two indices, percent correct and discrimination, are classical statistics that describe item data. Percent correct criteria at this phase of evaluation are stricter than those applied during calibration to identify items in need of additional field testing.

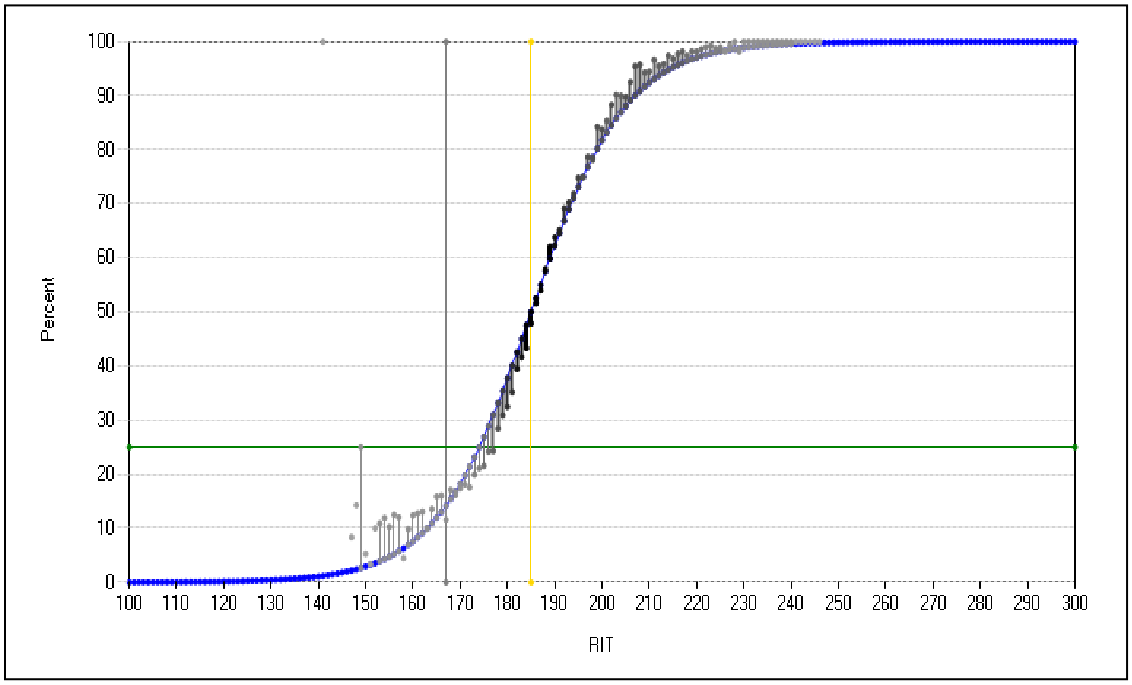
Table 5.4. Fit Index Descriptions and Criteria

Fit Index	Description	Criterion
Infit	Rasch weighted mean square fit statistic	< 1.09
Outfit	Rasch unweighted mean square fit statistic	< 1.09
MSF	Mean square fit	< 0.9
RMSE	Root mean squared error	< 1.0
Chi-square	Tests observed count correct versus expected count correct.	N/A
Std. Chi-square	Standardized chi-square statistic (Wilson & Hilferty, 1931)	< 1.0
r	Relationship between observed and expected values	> 0.75
Percent correct	Proportion of correct responses	$0.3 < p < 0.8$
Discrimination	Correlation between RIT score and item response	> 0.25

Graphic displays of item response functions are used to further evaluate items with borderline fit statistics. The item response function is a plot that shows the probability of a correct response to an item against the achievement levels of the students who responded to the item. When reviewing an item response display, the empirical item response function is plotted on the same grid as the theoretical function. When large discrepancies exist between the two curves, there is a lack of fit between the item and the scale. A more comprehensive understanding of item performance can be gained by reviewing the response functions. For example, if an item has a borderline chi-square value (indicating that performance on the item does not track well with increases in achievement), the item is flagged for revision or deletion.

Figure 5.1 presents the theoretical and empirical response functions for a Reading item with good fit to the Rasch model that was field tested by more than 4,000 students. The smooth curve shows the theoretical item response function from Equation 5.1 on p. 34, calibrated to the measurement scale based on all students responding. The vertical lines extending from the theoretical curve show the empirical proportion correct for the group of students with any final RIT score. Points not connected to the theoretical curve via a vertical line are based on small numbers of students (fewer than 10). The extent to which the empirical results deviate from the theoretical curve provides an index of item misfit. If the misfit is great, it might indicate that the item is flawed or that the model does not completely describe the item’s performance. The empirical results match the theoretical curve quite well, except in the extremes of the measurement range. However, in both the MAP Growth and the MAP Growth K–2 systems, items are targeted to the student’s performance, so it is rare that a student would see an item in the extremes of its measurement range. This item was approved for use in the item banks
.

Figure 5.1. Reading Item with Good Model Fit



5.4.2. Model of Man (MoM) Procedure

After the field test items calibrate through the item calibration engine, MoM is applied to the successfully calibrated items. The logistic regression model in MoM calculates the probabilities for each item that puts them into different status categories: “Auto Accept,” “Keep Field Test,” “Borderline Accept,” “Auto Reject,” and “Borderline Reject.” The MoM procedure was developed using a set of item calibration records containing 8,017 items across the four content areas (Reading, Language Usage, Mathematics, and Science) that were reviewed by two psychometricians over a 14-month period. The items were split into training and evaluation groups. Hauser et al. (2014) provided a detailed description of the MoM development process. They used the training group to build predictive models with a logistic regression approach with stepwise selection for each outcome, each for a content area, to identify the probability associated with decisions. The independent variables were the statistical indices calculated during the item calibration process. Experts’ item review decisions were used as a dependent variable. Statistically insignificant variables were dropped from the model.

5.4.3. Human Review Process

The human review process is conducted by psychometricians and content specialists. Once MoM provides the status categories to the successfully calibrated field test items, a visual review process is conducted by psychometricians who review the items by comparing the empirical item response function to the model-expected IRT. An item is flagged as “Auto Accepted” if its empirical and model item response functions are close across the RIT scale. If not, a psychometrician evaluates if the range of the differences is small. If the range is small and the total response count is larger than 1,000, the item is flagged as “Auto Accepted.” The item is flagged as “Keep Field Test” if the range is small and the total response count is less than 1,000. The “Auto Reject” flag is given to an item if the range of the differences is large. This visual process typically has three rounds of review involving at least two psychometricians:

1. In the first review, a psychometrician reviews all the “Borderline Reject,” “Borderline Accept,” “Auto Reject,” and “Auto Accept” items with item-total correlations above 0.10. The first reviewer also reviews most of the “Keep Field Test” items.
2. The second reviewer examines all the “Borderline Reject” and “Auto Reject” items accepted by the first reviewer and all the “Borderline Accept” and “Auto Accept” items rejected by the first reviewer.
3. The third review is only focused on the items that received different review decisions in the first two reviews.

Once psychometricians complete the visual review, the items flagged as “Auto Rejected” move to a post-calibration content review by content specialists who decide if the items could be revised or should be kept out of the MAP Growth item bank.

Chapter 6: Reporting

This chapter summarizes reporting information for MAP Growth assessments. For the Spanish Reading pilot, an overall RIT score and RIT scores by instructional area were provided. The learning continuum and learning statements were not available. Scores were not be used for any longitudinal reporting the pilot year. Only four of the traditional MAP Growth reports display non-growth events: Student Progress Report, Grade Report, Class Report, and Comprehensive Data File (CDF). However, growth events with a more extensive reporting suite were available beginning in Fall 2019. Currently all Spanish MAP Growth reports are in English. The only piece that is in Spanish is the instructional area label for Spanish Reading. For more details on the MAP Growth reports, please refer to the MAP Growth technical report.

6.1. MAP Growth Reports

Table 6.1 presents the required roles necessary to access the different report levels, and Table 6.2 summarizes the MAP Growth reports. In addition to these reports, the district assessment coordinator can use the Data Export Scheduler to export test results as CSV files to facilitate custom analysis and reporting.

Table 6.1. Required Roles for Report Access

Report Source	Required Role
Student-Level Reports	Instructor, Administrator, or District Assessment Coordinator
Class-Level Reports	Instructor, Administrator, or District Assessment Coordinator
District-Level Reports	Administrator or District Assessment Coordinator
Skills Checklist/Screening Reports	Instructor, Administrator, or District Assessment Coordinator
Learning Continuum	Instructor, Administrator, or District Assessment Coordinator

Table 6.2. Report Summary

Report Name	Description	Prior Data	Intended Audience
Student-Level Reports			
Student Profile	Brings together the data needed to advise each student and support their growth, including learning paths and growth goals.	All years prior	• Teacher • Instructional coach • Counselor • Student • Parent
Student Progress	Shows a student's overall progress from all past terms to the selected term to show the student's term-to-term growth.	All years prior	• Teacher • Instructional coach • Counselor • Student • Parent
Student Goal Setting Worksheet	Shows a student's test history and growth projections in the selected content areas for a specific period of time to discuss the student's goals and celebrate achievements.	Up to 2 years prior	• Teacher • Instructional coach • Counselor • Student • Parent

Report Name	Description	Prior Data	Intended Audience
Class-Level Reports			
Class	Shows class performance for a term, including norms status rankings, to analyze student needs.	1 year prior	• Instructional coach • Teacher
Achievement Status and Growth (ASG)	Shows three pictures of growth, all based on national norms: *projections* to set student growth goals, *summary* comparison of two terms to evaluate efforts, and an interactive *quadrant chart* to visualize growth comparisons.	Up to 2 years prior	• Instructional coach • Teacher • Counselor
Class Breakdown by RIT	Shows the academic diversity of a class across basic content areas to modify and focus the instruction for each student.	1 year prior	• Instructional coach • Teacher • Counselor
Class Breakdown by Goal	Shows the academic diversity for specific goals within a chosen content area to modify and focus the instruction for each student.	1 year prior	• Instructional coach • Teacher • Counselor
Class Breakdown by Projected Proficiency	Shows students' projected performance on state and college readiness assessments to adjust instruction for better student proficiency.	1 year prior	• Instructional coach • Teacher • Counselor • Principal
District-Level Reports			
District Summary	Summarizes RIT score test results for the current and all historical terms to inform district-level decisions and presentations.	All years prior	• Superintendent • Curriculum specialist • Instructional coach • Principal
Student Growth Summary	Shows aggregate growth in a district or school compared to the norms for similar schools to adjust instruction and use of materials.	All years prior	• Superintendent • Curriculum specialist • Instructional coach • Principal
Projected Proficiency Summary	Shows aggregated projected proficiency data to determine how a group of students is projected to perform on separate state and college readiness tests.	1 year prior	• Superintendent • Curriculum specialist • Instructional coach • Principal
Grade	Shows students' detailed and summary test data by grade for a selected term to set goals and adjust instruction.	1 year prior	• Principal • Counselor • Instructional coach
Grade Breakdown	Provides a single spreadsheet of student achievement (both subject and goal area) to flexibly group students from across the school. Unlike the Class Breakdown reports, this report has no limit on the number of students. File format is CSV.	1 year prior	• Principal • Counselor • Instructional coach

Report Name	Description	Prior Data	Intended Audience
Learning Continuum			
Class View	Shows students together with the skills and concepts they need to develop.	1 year prior	• Instructional coach • Teacher • Counselor
Test View	Shows skills and concepts for all RIT bands.	1 year prior	• Instructional coach • Teacher • Counselor

6.2. Learning Continuum

Every item in the NWEA item bank is associated with a learning statement, which is an instructionally relevant statement that describes the content the item is assessing. Learning statements are authored and assigned to items by NWEA content specialists. A content specialist will review an item—its intent, target, and existing standard alignments—and select or write a learning statement that captures the content of the item (without describing the item in detail). Learning statements allow NWEA to describe the contents of a MAP Growth assessment without exposing the items themselves. Because learning statements are assigned to items, they have indirect relationships to standard statements, RIT values, and other data points via the items. These relationships among learning statements, standards, and RIT values form the basis of the learning continuum. The Spanish learning continuum also has learning statements that are unique to Spanish. These are largely in the foundational skills, grammar, mechanics, and conventions areas. Spanish shares learning statements with ELA in instances when the targeted constructs are the same. Learning statements on the Spanish learning continuum are in English.

The learning continuum, designed for classroom use, translates MAP Growth scores to learning statements that show what students performing at a given RIT level on MAP Growth assessments are typically ready to learn to allow teachers to set student goals and tailor instruction to student needs. The learning continuum identifies skills and concepts each student is ready to learn by showing relationships among standards, learning statements, and the student’s RIT score. This helps educators bridge the gap between MAP Growth data and standards and/or intervention.

Chapter 7: Reliability

Reliability refers to the consistency of scores obtained from the assessment. It reflects the absence of random measurement error. When the measurement error is small, reliability is large, and vice versa. Increasing reliability by minimizing error is an important goal for any test. Different sources of measurement error affect scores. The effect of each particular source of error has a corresponding reliability coefficient that describes the influence of that source on scores. One source of measurement error is time, or the instability of a construct over time, as measured by test-retest reliability. If this source of error is low, the test-retest reliability coefficient will be high. Another source of measurement error is the items selected for a test. Internal consistency, or marginal reliability, will be high if measurement error due to items is low.

It is important to report multiple reliability coefficients to describe the influence of different sources of error. Therefore, the reliability of the Spanish MAP Growth Reading assessments was examined in the following ways:

- **Test-retest reliability** that demonstrates the consistency of MAP Growth assessments across time by administering it to a group of students two times separated by a reasonable period of time. The question being answered with this type of reliability is “To what extent does the test administered to the same students twice yield the same results from one administration to the next?”

- **Marginal reliability** that examines a test’s consistency across items. The question being answered with this type of reliability is “To what extent do items in the test measure the test’s construct(s) in a consistent manner?”

- **Score precision** based on the standard error of measurement (SEM) of MAP Growth scores

Data included in these analyses were from the Fall 2018, Winter 2019, and Spring 2019 administrations of the Spanish MAP Growth Reading pilot data.

7.1. Test-Retest Reliability

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 goal structure, 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 where several months separate the two administrations instead of the typical two or three weeks. The second test (or retest) is not the same test. Rather, it is one that is comparable to the first by its content and structure, differing only in the difficulty level. In other words, test-retest with alternate forms (Crocker & Algina, 1986) describes the influence of two sources of measurement error: time and item selection.

Specifically, 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., e.g., Fall 2018 and Winter 2019; Winter 2019 and Spring 2019). Table 7.1 presents test-retest reliability results by grade. The grade level is based on students' actual grade levels. Coefficients range from 0.50 at Grade K to 0.83 at Grade 5. There is no test-retest reliability for Grades 6–8 because only one test administration was taken by these grades.

Table 7.1. Test-Retest with Alternate Forms Reliability by Grade

	Fall 2018 – Winter 2019		Fall 2018 – Spring 2019		Winter 2019 – Spring 2019	
Grade	N	Reliability	N	Reliability	N	Reliability
K	4,143	0.580	4,759	0.503	5,846	0.615
1	4,923	0.743	5,756	0.703	7,164	0.777
2	5,215	0.757	5,322	0.744	7,039	0.774
3	3,921	0.786	4,178	0.777	5,611	0.794
4	3,315	0.802	3,723	0.781	4,948	0.806
5	2,709	0.831	2,721	0.824	4,339	0.819

7.2. Marginal Reliability (Internal Consistency)

Internal consistency measures how well the items on a test that reflect the same construct yield similar results. Determining the internal consistency of MAP Growth tests is challenging because traditional methods depend on all test takers taking a common test consisting of the same items. Application of these methods to adaptive tests is statistically cumbersome and inaccurate. Fortunately, an equally valid alternative is available in the marginal reliability coefficient (Samejima, 1977, 1994) that incorporates measurement error as a function of the test score. In effect, it is the result of combining measurement error estimated at different points on the achievement scale into a single index. This method of calculating internal consistency, ρ_θ, yields results that are nearly identical to coefficient alpha when both methods are applied to the same fixed-form tests. The approach taken for MAP Growth was suggested by Wright (1999) and is given by:

$$\rho_\theta = \frac{\sigma_\theta^2 - M_{S_\theta^2}}{\sigma_\theta^2} \qquad (7.1)$$

where σ_θ^2 is the observed variance of the achievement estimates, θ, (the RIT score) and $M_{S_\theta^2}$ is the observed mean of the score's conditional error variances at each value of θ. Tests are considered of sound reliability when their marginal reliability coefficients range from 0.80 and above.

Table 7.2 presents the marginal reliabilities of RIT scores by grade. The overall marginal reliabilities for all grades are in the 0.90s, which suggests that the Spanish MAP Growth Reading tests have high internal consistency.

Table 7.2. Marginal Reliability by Grade

Grade	N	Reliability	Mean SEM
K	14,765	0.926	3.7
1	17,866	0.948	3.5
2	17,573	0.945	3.6
3	13,709	0.955	3.6
4	11,987	0.959	3.5
5	9,770	0.961	3.5
6	2,371	0.942	3.6
7	1,999	0.950	3.6
8	1,628	0.956	3.7

7.3. Score Precision

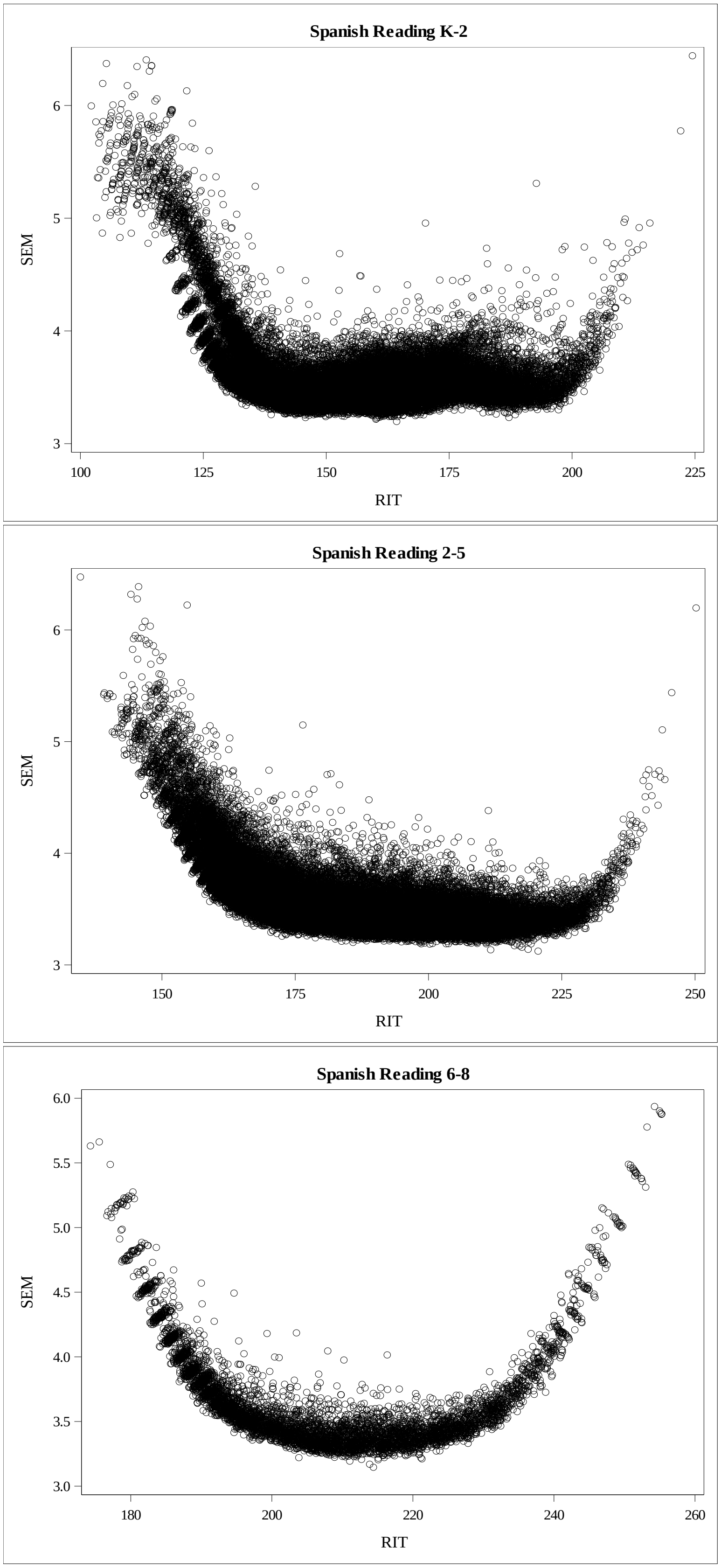
Score precision is measured by the standard error of measurement (SEM), a function of the relationship among item parameters, the ability of the student, and the number of items administered. SEM is related to reliability in that it estimates how repeated measures of a student on the same assessment tend to be distributed around their “true” score. The SEM is the inverse of the square root of test information. Score precision is best when students are given items closely matched to their abilities. Lower values of SEM indicate greater precision in the score. With greater score precision across a broad range of ability, several benefits follow:

- Differences between similar students become more apparent. Because of the direct mathematical relationship between test information and SEM, a lower SEM indicates greater test information, so the level of test information across a group of students from even a wide grade span should be comparable across the achievement range.
- When change in student scores from one test occasion to another is of interest, measurement errors accrue with each test occasion. The greater the precision of individual scores, the greater the likelihood of drawing reliable conclusions about changes in student status over time.
- Classification accuracy will be improved as the level of score precision is increased.

The MAP Growth adaptive test algorithm selects the best items for each student, producing a significantly lower SEM than fixed-form tests. MAP Growth tests yield ability estimates with SEMs that are less than .30 of a typical large sample standard deviation (Kingsbury & Hauser, 2004). Standard errors vary minimally across more than 90% of the achievement range of a grade level. This makes MAP Growth scores well suited for use in growth models and other statistical procedures that assume additive measures.

Figure 7.1 presents the levels of SEM across the operational RIT range for Spanish MAP Growth tests by grade band. Each figure has a noticeable fluctuation in SEMs at the very low and very high end of the RIT score distributions. All mean SEMs are below 4.5 RITs except at the very low and high levels of the RIT score distributions, which is to be expected. This consistency in MAP Growth SEMs across the RIT ranges of interest is particularly important when student change in performance is to be evaluated. Because Spanish MAP Growth is used to monitor students’ progress over years, it is important that Spanish MAP Growth has similarly low SEMs across the RIT score range so that students at different ability levels are measured equally precisely.

Figure 7.1. SEM of RIT Scores



Spanish Reading K-2

Spanish Reading 2-5

Spanish Reading 6-8

Chapter 8: Validity

Validity is defined as the "the degree to which evidence and theory support the interpretations of test scores for proposed uses. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (AERA et al., 2014, p. 11). It is not a quantifiable property but an ongoing process, beginning at initial conceptualization of the construct, continuing throughout the entire testing process, and extending into the interpretation and use of test sores. Validity evidence for Spanish MAP Growth Reading involves multiple sources including test content, internal structure, and relations to other variables.

8.1. Evidence Based on Test Content

Chapter 2 describes test content and alignment to standards, and Chapter 3 describes item development procedures. This information is important for providing evidence that the Spanish MAP Growth Reading assessments were developed to coincide with their intended purposes.

8.2. Evidence Based on Internal Structure

The degree to which adapted versions of tests are equivalent across languages or language groups is an important issue in considering the validity of tests. According to the *Standards* (AERA et al., 2014, p. 68–69), the Guidelines for Translating and Adapting Tests (ITC, 2016), and numerous researchers (e.g., Hambleton, 2005; Sireci, 2011; Van de Vijver & Poortinga, 2005), empirical evidence such as test construct equivalency and DIF across language groups must be provided to support the validity of inferences derived from cross-lingual assessments. The assumptions for using the Spanish MAP Growth Reading test compared to the English MAP Growth test are as follows:

- For Spanish speakers with different language backgrounds, the Spanish MAP Growth Reading test measures the same Spanish reading construct.
- For dual-language English and Spanish speakers, the Spanish and English MAP Growth Reading tests measure similar but different constructs (about 1/3 of the Spanish MAP Growth Reading items do not exist in the English version).

The evidence of equivalence across languages and language backgrounds at the individual item level has been provided with the DIF results (Table 8.4). To provide evidence of the equivalence of the internal structure of the Spanish MAP Growth Reading test across Spanish speakers' language backgrounds, this section focuses on goodness of fit summary indices using the confirmatory factor analysis (CFA) method, which is a multivariate statistical procedure used to test how well the measured variables represent the number of constructs. The purpose of CFA in this instance is to confirm or reject the Spanish MAP Growth Reading measurement theory.

8.2.1. Goodness-of-Fit Indices

Model assumptions include congeneric, tau-equivalent, and parallel, as described below (Byrne et al., 1989; Feldt & Brennan, 1989; Jöreskog & Sörbom, 1979; Loehlin, 2004). The equivalence of the factor loading and the variance of three factor models was tested by placing different constraints (equal loading or variance) on the two compared models (e.g., congeneric compared to tau-equivalent, tau-equivalent compared to parallel).

- Congeneric: The least restrictive of the three models in which both factor loadings and error variances are free estimated (without constraints). This model assumes that each measured variable (goal score) measures the same latent variable (achievement) with different degrees of precision and different errors.
- Tau-equivalent: A less restrictive model that is identical to the parallel model except error variances are free. This model implies that each measured variable (goal score) measures the same latent variable (achievement) with the same degree of precision, but with possible different errors.
- Parallel: This is the most restrictive model of the three. The two tests are psychometrically parallel if they share an equal amount of factor loadings and the error variances of observed variables or they are fixed to be equal. This means that each measured variable (instructional area score) measures the same latent variable (achievement) with the same degree of precision and the same scale (Raykov, 1997a, 1997b).

Goodness-of-fit describes how well a statistical model fits a set of observations. Measures of model fit typically summarize the discrepancy between observed values and the values expected under the model in investigation. The following well-known goodness-of-fit indices were used to evaluate model fit for the Spanish MAP Growth Reading test:

- Absolute indexes that include chi-square χ^2, unadjusted goodness-of-fit indices (GFI), adjusted GFI (AGFI), and standardized root mean square residual (SRMR)
- Incremental indexes Bentler-Bonett normal fit index (NFI)
- The parsimony index that includes the root mean square error of approximation (RMSEA)

Although the χ^2 value provides the basis of comparison with the previously fitted model, χit is not considered as the best practice because it is sample size dependent. A significant χ^2 does not necessarily indicate a departure from invariance when the sample size is large (e.g., 5,000). Hu and Bentler (1999) recommended using combinations of goodness-of-fit indices to obtain a robust evaluation of data-model fit in structural equation modeling. They recommend the following cutoff criterion values of good model fit: GFI, AGFI, NFI > 0.95, RMSEA < 0.06, and SRMR < 0.08.

However, many researchers (e.g., Marsh et al., 2005) demonstrated that these criteria are too restrictive. For example, some researchers believe that these cutoff values are too rigorous and may have limited generalizability to the levels of misspecification experienced in typical practice (Beauducel & Wittmann, 2005; Fan & Sivo, 2005; Marsh et al., 2005; Yuan, 2005). Therefore, a “good enough” or “rough guideline” approach for absolute fit indices and incremental fit indices (such as GFI and NFI) have been commonly accepted (Lance et al., 2006). Under the relaxed criteria, cutoff values should be above 0.90, and values below 0.10 are usually considered adequate for fit indices based on residuals matrix (such as RMSEA and SRMR).

8.2.2. Results

To evaluate model fit for the Spanish MAP Growth Reading assessment, Table 8.1 presents the summary of the goodness-of-fit indices for testing congeneric, tau-equivalent, and parallel equivalences of the Spanish tests across different language background groups based on both Hu and Bentler’s criteria and the relaxed criteria (only CCSS data were used for the goodness-of-fit summary indices). The Akaike’s information criterion (AIC) statistic is also reported to test the data-model fit to further verify the model selection. Degree of freedom (df) is also provided. Blank cells indicate no results because of a low n-count. All analyses were conducted using SAS 9.4.

All the indices express results across different models in a nested series of tests. First, all χ^2 values are statistically significant at 0.01 α-level, which could mean that all models could be rejected. However, χ^2 is a questionable indicator when the sample size is large, so a significant value of χ^2 in this case does not necessarily indicate it is necessary to reject the hypothesized model since the sample size is large. A chi-square test correlates with sample size and would detect even minimal differences between the hypothesized model and the data (Bollen & Long, 1993; Browne & Cudeck, 1993). For all models tested, the fit indices of GFI, AGFI, NFI, RMSEA, and SRMR all exceed the relaxed criteria. For most congeneric, tau-equivalent, and parallel models, the fit indices of GFI, AGFI, NFI, RMSEA, and SRMR exceed Hu and Bentler’s criteria indicate good data model fit.

Overall, the results indicate that the constructs measured by the Spanish MAP Growth Reading tests across language background groups are at least tau-equivalent and most of them are parallel equivalent. The major implication of these results is that Spanish MAP Growth Reading tests can be used for students with different language backgrounds who receive different classroom instruction.

Table 8.1. Goodness-of-Fit Summary Indices

				Goodness-of-Fit Indices**						
Compared Groups*	**Model**	**N-Count**	**df**	χ^2	**GFI**	**AGFI**	**NFI**	**RMSEA**	**SRMR**	**AIC**
Grades K–2										
E_ES/S_ES	Congeneric	14,505	10	240.00	0.99	0.98	1.00	0.06	0.01	260.00
	Tau-equivalent	14,505	14	303.14	0.99	0.99	0.99	0.05	0.05	315.14
	Parallel	14,505	15	318.18	0.99	0.99	0.99	0.05	0.05	328.18
E_ES/S_E	Congeneric	1,509	10	73.96	0.98	0.98	0.98	0.05	0.01	93.02
	Tau-equivalent	1,509	14	87.91	0.97	0.96	0.98	0.08	0.14	99.91
	Parallel	1,509	15	88.21	0.97	0.96	0.98	0.08	0.15	98.21
E_ES/S_ES	Congeneric	5,116	10	129.01	0.99	0.96	0.99	0.07	0.01	149.01
	Tau-equivalent	5,116	14	174.89	0.98	0.98	0.99	0.06	0.04	186.89
	Parallel	5,116	15	204.47	0.98	0.98	0.99	0.07	0.04	214.47
E_ES/S_S	Congeneric	1,732	10	75.34	0.98	0.98	0.99	0.08	0.02	95.33
	Tau-equivalent	1,732	14	88.44	0.98	0.98	0.99	0.07	0.04	100.44
	Parallel	1,732	15	118.16	0.96	0.95	0.98	0.07	0.04	214.47
ES_ES/S_ES	Congeneric	4,028	10	81.13	0.99	0.98	0.99	0.05	0.01	101.13
	Tau-equivalent	4,028	14	99.99	0.99	0.98	0.99	0.08	0.08	111.99
	Parallel	4,028	15	105.72	0.99	0.98	0.98	0.09	0.10	115.72

				Goodness-of-Fit Indices**						
Compared Groups*	Model	N-Count	df	χ^2	GFI	AGFI	NFI	RMSEA	SRMR	AIC
Grades 2–5										
E_ES/S_ES	Congeneric	15,547	18	359.87	0.99	0.99	0.99	0.05	0.01	383.87
	Tau-equivalent	15,547	23	406.51	0.99	0.99	0.99	0.05	0.03	420.51
	Parallel	15,547	24	585.58	0.99	0.99	0.99	0.05	0.04	597.58
E_ES/S_E	Congeneric	1,946	18	197.87	0.96	0.93	0.97	0.10	0.03	221.87
	Tau-equivalent	1,946	23	204.82	0.96	0.95	0.97	0.09	0.03	218.82
	Parallel	1,946	24	226.17	0.96	0.95	0.97	0.09	0.03	238.17
E_ES/S_ES	Congeneric	7,378	18	277.80	0.99	0.99	0.99	0.06	0.02	301.80
	Tau-equivalent	7,378	23	319.97	0.98	0.98	0.99	0.06	0.06	238.17
	Parallel	7,378	24	625.32	0.97	0.96	0.99	0.08	0.07	637.32
E_ES/S_S	Congeneric	2,492	18	178.02	0.97	0.96	0.98	0.08	0.03	202.02
	Tau-equivalent	2,492	23	299.30	0.96	0.96	0.97	0.10	0.22	313.30
	Parallel	2,492	24	517.52	0.90	0.87	0.95	0.13	0.26	529.52
ES_ES/S_ES	Congeneric	6,413	18	190.75	0.97	0.98	0.98	0.08	0.03	214.75
	Tau-equivalent	6,413	23	240.26	0.98	0.98	0.97	0.05	0.06	254.26
	Parallel	6,413	24	259.69	0.98	0.98	0.99	0.05	0.07	271.67
Grades 6–8										
E_ES/S_ES	Congeneric	3,574	18	279.26	0.97	0.95	0.97	0.09	0.04	303.26
	Tau-equivalent	3,574	23	292.78	0.97	0.95	0.97	0.08	0.06	306.48
	Parallel	3,574	24	454.58	0.96	0.93	0.95	0.10	0.05	446.58
E_ES/S_E	Congeneric	–	–	–	–	–	–	–	–	–
	Tau-equivalent	–	–	–	–	–	–	–	–	–
	Parallel	–	–	–	–	–	–	–	–	–
E_ES/S_ES	Congeneric	2205	18	188.43	0.97	0.97	0.97	0.09	0.04	212.43
	Tau-equivalent	2,205	23	201.88	0.97	0.97	0.97	0.08	0.08	215.89
	Parallel	2,205	24	334.18	0.95	0.95	0.94	0.10	0.06	346.18
E_ES/S_S	Congeneric	–	–	–	–	–	–	–	–	–
	Tau-equivalent	–	–	–	–	–	–	–	–	–
	Parallel	–	–	–	–	–	–	–	–	–
ES_ES/S_ES	Congeneric	2,270	18	312.42	0.95	0.95	0.95	0.12	0.05	336.43
	Tau-equivalent	2,270	23	321.12	0.95	0.95	0.95	0.10	0.08	335.12
	Parallel	2,270	24	336.12	0.94	0.94	0.95	0.11	0.09	348.12

*E_ES = native English speakers with English and Spanish instruction. S_All = all Spanish speakers regardless of instruction. S_E = native Spanish speakers with English instruction. S_ES = native Spanish speakers with English and Spanish instruction. S_S = native Spanish speakers with Spanish instruction. ES_ES = native English and Spanish speakers with either English or Spanish instruction or both.

**GFI = unadjusted goodness-of-fit indices. AGFI = adjusted goodness-of-fit indices. NFI = normal fit index. RMSEA = root mean square error of measurement. SRMR = standardized mean root mean square residual. AIC = Akaike's information criterion.

8.3. Differential Item Functioning (DIF)

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

When students with the same ability from two different groups of interest have different probabilities of correctly answering an item, the item is said to exhibit DIF, a statistical characteristic of an item that shows the extent to which the item might be measuring different ability for different student subgroups. DIF indicates a violation of a major assumption of the Rasch model, and it signals potential for a lack of fairness at the item level. The presence of DIF in an item suggests that the item is functioning unexpectedly regarding the groups included in the comparison. The cause of the unexpected functioning is not revealed in a DIF analysis. It may be that item content is inadvertently providing an advantage or disadvantage to members of one of the two groups. Content experts who have special knowledge of the groups involved are often in a good position to identify a cause of this type. DIF may also result from differential instruction closely associated with group membership.

The Mantel-Haenszel (MH) procedure (1959) is the most cited and studied method for detecting DIF. It stratifies examinees by a composite test score, compares the item performance of reference and focal group members in each strata, and then pools this comparison over all strata. The MH procedure is easy to implement and is featured in most statistical software. NWEA applied the MH method to assess DIF for the MAP Growth Spanish Reading item pool.

The results are categorized based on the Educational Testing Service (ETS)'s method of classifying DIF (Zwick, 2012). Table 8.2 presents the criteria for each level of classification. This method allows items exhibiting negligible DIF (Category A) to be differentiated from those exhibiting moderate DIF (Category B) and strong DIF (Category C). Categories B and C have a further breakdown as "+" (DIF is in favor of the focal group) or "-" (DIF is in favor of the reference group). All items exhibiting moderate (Category B) DIF are subjected to an extra review by content specialists to identify the source of DIF. For each item, these specialists decide the following:

- Remove the item from item bank
- Revise the item and re-submit it for field testing
- Retain the item without modification

Items exhibiting strong DIF (Category C) may or may not be removed during pilot field test review and final operational item review.

Table 8.2. DIF Categories

DIF Category	Level of DIF	Definition
A	Negligible	• Absolute value of the Mantel-Haenszel delta difference (MH D-DIF) is not significantly different from 0 or is less than one.
B	Moderate	• Absolute value of the MH D-DIF is significantly different from 0 but not from one, and is at least 1; or • Absolute value of the MH D-DIF is significantly different from 1, but less than 1.5. • Positive values are classified as “B+” and negative values as “B-“.
C	Strong	• Absolute value of the MH D-DIF is significantly different from 1, and is at least 1.5; and • Absolute value of the MH D-DIF is larger than 1.96 times the standard error of MH D-DIF. • Positive values are classified as “C+” and negative values are “C-“.

Two kinds of data for the DIF analyses taken from of the Spanish pilot include bilingual and monolingual data. The bilingual data were used for detecting DIF for transadapted items, and monolingual data were used to detect DIF for the newly developed Spanish items used only on the Spanish tests. Purposes of this DIF study are as follows:

- To examine the quality of the transadapted items and the impact of language effect for the bilingual data using both English and Spanish responses
- To examine the language background impact on the Spanish items from the monolingual data.

Although all students took the Spanish version, they may come from four different language background groups:

- Native English speakers who received instruction in both English and Spanish
- Native Spanish speakers who received instruction in just English
- Native Spanish speakers who received instruction in just Spanish
- Native Spanish speakers who received instruction in both English and Spanish

Not all schools provided this information because it was voluntary-based, but about half of the students had this information in the pilot data. The characteristics of Spanish speakers provide vital information on the effect of test scores, which is important for making sure the test scores are useful based on their intended purposes for the target population. For example, the same Spanish MAP Growth Reading score may reveal additional information on a student’s ability to learn and develop if the language background information is known.

Table 8.3 presents the sample size and final number of operational items included in the DIF study. The Native English/Bilingual group refers to native English with instruction in English and Spanish vs. native Spanish with instruction in either English or Spanish, or both. The Native Spanish/Bilingual group refers to native Spanish with instruction in Spanish vs. native Spanish or native English with instruction in either English or Spanish, or both.

Table 8.3. Number of Items in DIF Study and Descriptive Statistics of DIF Student Sample

DIF Study Group (Focal/Reference)	Data	#Items*	Student Sample			
			Mean	SD	Min.	Max.
Language (English/Spanish)	Bilingual	604	1308.57	903.87	175	7115
Gender (Female/Male)	Monolingual	919	1563.05	1318.06	163	8213
Native English/Bilingual	Monolingual	916	1589.98	1330.86	163	8213
Native Spanish/Bilingual	Monolingual	919	1548.56	1315.44	163	8213

*Among the 919 operational items, 604 of them are transadapted items from the English version.

Table 8.4 presents the number of items and percentage of items exhibiting DIF by language, gender, or language background for the Spanish MAP Growth Reading pilot. Only one transadapted item shows Category C DIF across languages. For DIF related to gender and language background, the percentage of Category C DIF for both are about 1%. The Native English/Bilingual group had the highest percentage of items. Overall, these DIF results show the following three patterns:

- Most items are classified as A.
- Highest percentage of C DIF is the Native English/Bilingual group (6.66%).
- C DIF is rare for the remaining DIF study groups (~1%).

Table 8.4. DIF Results

	DIF Results by DIF Study Group							
	English/Spanish		Female/Male		Native English/Bilingual		Native Spanish/Bilingual	
ETS Class	#Items	%	#Items	%	#Items	%	#Items	%
A	594	98.34	875	95.21	634	69.21	875	95.21
B+	7	1.16	20	2.18	56	6.11	20	2.18
B-	2	0.33	15	1.63	139	15.17	15	1.63
C+	–	–	4	0.44	26	2.84	4	0.44
C-	1	0.17	5	0.54	61	6.66	5	0.54

Chapter 9: Pilot User Norms

Apart from interpretations of performance and growth regarding content, how students performed or grew compared to an appropriate reference peer group (provided by norms) is important information for individualizing instruction, setting achievement goals for students or entire schools, understanding achievement patterns, and evaluating student performance. As of the July 2019 release of the Spanish MAP Growth Reading test, pilot user norms were made available for Spanish Reading within MAP Growth reports based on the pilot year of test data. Although they are drawn from a limited pool of test events and not a nationally representative like the general MAP Growth norms, they provide basic contextual information about student performance in the fall and spring and growth between fall and spring on the Spanish MAP Growth Reading assessments. NWEA intends to refresh the Spanish MAP Growth Reading user norms data for Fall 2020 based on available testing data pool from the 2019–2020 school year.

9.1. Pilot User Norms vs. National Norms

The major difference between pilot user norms and the general MAP Growth Reading norms is that the Spanish Reading pilot user norms are drawn from a limited pool of test events during the pilot year and are not nationally representative. Nationally representative norms can be calculated when the number of responses is very large and the dataset includes the following:

- Responses from all or most U.S. states
- A distribution across urban, rural, suburban districts and schools
- Large and small districts and schools
- High and low socioeconomic groups

For nationally representative norms, the sample mirrors the national population of students as a whole to permit comparisons of individual or group performance to students across the nation. For Spanish MAP Growth Reading, student performance is compared to other students in the same grade who took the assessments during the pilot year. The sample for Spanish Reading does not meet the criteria listed above and therefore cannot support nationally representative norms. Table 9.1 presents the key differences between the norms for both the Spanish and English versions of MAP Growth.

Table 9.1. Pilot Spanish User Norms vs. English National Norms

	Spanish Norms	English Norms
Within-year growth norms	Fall-to-spring for Grades K–5	• Fall-to-winter • Winter-to-spring • Fall-to-spring
Between-year growth norms	Not available	• Fall-to-fall • Winter-to-winter • Spring-to-spring
Achievement norms	• Fall and spring norms that are specific to a student's grade for K–5 • Spring norms that are specific to a student's grade for 6–8	Fall, winter, and spring norms specific to a student's grade
Instructional weeks	Not adjusted for instructional weeks as configured by each partner	Adjusted for instructional weeks as configured by each partner
School norms	Not available	Available

9.2. Estimating the Pilot User Norms

Unlike the nationally representative English MAP Growth Reading norms that employed a three-level hierarchal linear model (HLM) to reflect the nesting of repeated observations of students within schools for modeling growth (Thum & Hauser, 2015), the Spanish MAP Growth Reading pilot user norms employed a multivariate true score model (MTSM) to reflect the nesting of repeated observations of students within schools for modeling growth (Thum & He, 2018).

In contrast to the HLM method used in the regular MAP Growth norms, MTSM used for the Spanish pilot does not require three-years of growth data or the weights based on the School Challenge Index (SCI), which is a measure of how U.S. public schools compare in terms of challenges and opportunities they operate under. Inferences based on MTSM rely on the reasonableness of the joint normality assumption of score components for their validity. Normality was examined from different perspectives such as quantile-quantile (Q-Q) plots, cumulative distribution function (CDF) curves for RIT scores, and residuals from model estimation.

9.3. Norms Reference Groups

The Spanish MAP Growth Reading norms were created using the Grades K–8 pilot sample collected from August 2018 to July 2019. Table 9.2 presents the mean and standard deviation (SD) of RIT test scores for grades with valid test events. Students in Grades 6–8 only took spring tests. As expected, average test scores increase as grades increase, and average test scores also increase as students grew from fall to spring.

Table 9.2. Summary Descriptive Statistics of Sample Test Scores

Grade		Fall	Spring
K	Mean	136.18	148.36
	SD	8.50	13.54
	N	4,140	5,862
1	Mean	150.26	163.51
	SD	13.04	14.95
	N	4,933	7,171
2	Mean	167.28	176.60
	SD	13.49	15.74
	N	5,217	7,038
3	Mean	178.57	187.81
	SD	14.88	17.20
	N	3,918	5,609
4	Mean	186.64	195.33
	SD	15.74	17.49
	N	3,318	4,947
5	Mean	192.77	200.16
	SD	17.15	17.70
	N	2,709	4,340
6	Mean	–	203.63
	SD	–	14.50
	N	–	2,232

Grade		Fall	Spring
7	Mean	–	209.47
	SD	–	16.14
	N	–	1,946
8	Mean	–	211.87
	SD	–	17.65
	N	–	1,604

9.4. Achievement Status and Growth Norms

Norms provide the relative performance of students in a specified population. In achievement status norms, a student's performance on the Spanish MAP Growth Reading assessments, expressed as a RIT score, is associated with a percentile ranking that shows how well the student performed compared to students in the norming group. The relative evaluation of a student's growth from one period to another (e.g., from fall to spring) is provided by growth norms.

Appendix A presents the achievement status and growth norms, and Table 9.3 presents a snapshot of the achievement status and fall-to-spring growth norms and their associated percentiles for the Grade 1 Spanish MAP Growth Reading test for illustrative purposes, as well as the expected fall-to-spring gain and SD of predicted growth score. For ease of presentation, not every possible percentile is provided in this table. The red numbers indicate spring achievement status norms and their corresponding percentiles. The blue numbers indicate the growth percentiles associated with fall-to-spring growth scores. For example, the 55th achievement percentile scores for fall and spring are 151 and 165, respectively. The expected fall-to-spring gain for a student who starts in the fall at the 55th percentile score of 151 is 13.6 with an associated SD of growth of 9.5. This indicates that students who perform at the 55th percentile in the fall test tend to present an average growth of 13.6 RITs in the spring.

Table 9.3 allows readers to normatively evaluate the actual gain a student may have made from fall to spring. For example, if a student who scores 151 in the fall (55th percentile) obtains a score of 169 in the spring (65th percentile), this student has improved 18 RITs (167-151=18) from fall to spring. Locating the intersection in the table, corresponding to the row where the achievement percentile is 55 and to the column where the spring score percentile is 65, the 18 fall-to-spring RIT gain puts this student at the 67th percentile in the specific growth scale.

Table 9.3. Status and Growth Norms for Spanish MAP Growth Reading Test of Grade 1

	Achievement Status Norms		Fall-to-Spring Conditional Growth Norms*		Spring Quantile and RIT																		
					5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)	140	145	149	151	154	156	158	160	162	164	165	167	169	171	173	176	179	182	188
5	129	140	16.5	9.5	27	48	62	73	81	86	90	93	96	97	98	99	99	99	99	99	99	99	99
10	134	145	15.9	9.5	15	32	46	58	67	75	81	86	90	93	95	97	98	99	99	99	99	99	99
15	137	149	15.5	9.5	9	22	35	46	57	65	73	79	84	88	92	94	96	98	99	99	99	99	99
20	139	151	15.2	9.5	6	16	27	38	48	57	65	72	78	83	88	91	94	96	98	99	99	99	99
25	141	154	14.9	9.5	4	12	21	31	40	49	57	65	72	78	83	88	91	94	96	98	99	99	99
30	143	156	14.6	9.5	3	9	17	25	34	42	51	59	66	73	79	84	88	92	95	97	99	99	99
35	145	158	14.4	9.5	2	7	13	20	28	36	44	52	60	67	74	80	85	89	93	96	98	99	99
40	147	160	14.2	9.5	1	5	10	16	23	31	38	46	54	61	69	75	81	86	91	94	97	99	99
45	148	162	14.0	9.5	1	4	8	13	19	26	33	41	48	56	63	70	77	83	88	92	96	98	99
50	150	164	13.8	9.5	1	3	6	10	15	21	28	35	42	50	58	65	72	79	85	90	94	97	99
55	151	165	13.6	9.5	1	2	4	8	12	17	23	30	37	44	52	60	67	74	81	87	92	97	99
60	153	167	13.3	9.5	1	1	3	6	9	14	19	25	32	39	46	54	62	69	77	84	90	95	99
65	155	169	13.1	9.5	1	1	2	4	7	11	15	20	26	33	40	48	56	64	72	80	87	93	98
70	156	171	12.9	9.5	1	1	2	3	5	8	12	16	21	27	34	41	49	58	67	75	84	91	97
75	158	173	12.6	9.5	1	1	1	2	4	6	9	12	17	22	28	35	43	51	60	69	79	88	96
80	160	176	12.4	9.5	1	1	1	1	2	4	6	9	12	17	22	28	35	43	52	62	73	84	94
85	163	179	12.0	9.5	1	1	1	1	1	2	4	6	8	12	16	21	27	35	44	54	65	78	91
90	166	182	11.6	9.5	1	1	1	1	1	1	2	3	5	7	10	14	19	25	33	43	54	68	85
95	171	188	11.0	9.5	1	1	1	1	1	1	1	1	2	3	5	7	10	14	20	27	38	53	73

*E(S-F|F) = expected fall-to-spring growth given a fall score. SD(S-F|F) = standard deviation of fall-to-spring growth given a fall score.

Since these Spanish MAP Growth Reading pilot norms are based on pilot data and given the available evidence employed to construct these norms, users should exercise caution about the limited generalizability of the inferences supported by the results presented in these norms. For example, instructional decisions that rely on inferences about the normative performance of students are likely to be less precise. Similarly, the lower precision in these norms should be factored into secondary or derived uses of student normative scores such as teacher or school accountability. While NWEA will continue to improve these norms as more data become available over time, these norms should offer a first attempt to schools, teachers, or parents to interpret and understand how students are performing at a point in time and over the course of the year in Spanish MAP Growth Reading.

References

Achieve. (2018). *A framework to evaluate cognitive complexity in mathematics assessments*. www.achieve.org/files/Cognitive%20Complexity%20Mathematics%20Assessment_FINAL_0.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.

Anderson, L. W., & Krathwohl, D. R. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*(1), 41–75. doi.org/10.1207/s15328007sem1201_3

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. SAGE Publications.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). SAGE Publications.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(30, 456–466.

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*.

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

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12*(3), 343–367. doi.org/10.1207/s15328007sem1203_1

Feldt, L., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.,), *Educational measurement* (3rd ed., pp. 105–146). MacMillan.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Lawrence Erlbaum Publishers.

Hauser, C., Thum, Y. M., He, W., & Ma, L. (2014). Using a model of analysts' judgments to augment an item calibration process. *Educational and Psychological Measurement, 75*(5), 826–849.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi.org/10.1080/10705519909540118

Ingebo, G. S. (1997). *Probability in the measure of achievement*. MESA Press.

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Jiban, C. (2017). *MAP Growth Reading and Language Usage literature review*. NWEA.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Abt Books.

Kingsbury G. G., & Hauser, C. (2004, April). *Computerized adaptive testing and No Child Left Behind.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Diego, CA.

Kingsbury, G. G., & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359–375.

Kingsbury, G. G., & Zara, A. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*(3), 241–261.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). *Goodness of fit in structural equation models.* In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Lawrence Erlbaum Associates.

National Center for Education Statistics (NCES). (2019). *Fast facts: English language learners*. nces.ed.gov/fastfacts/display.asp?id=96

National Governors Association Center for Best Practices & Council of Chief State School Officers (CCSSO). (2012). *Common core state standards Spanish language version*. commoncore-espanol.sdcoe.net/

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. MESA Press.

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173–184. doi:10.1177/01466216970212006

Raykov, T. (1997b). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*(4), 375–385. doi:10.1177/014662169802200407

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*(3)*,* 233–247.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*(3), 229–244.

Sireci, S. G. (2011). Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 216–240). Oxford University Press.

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 (NCEO). education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. NWEA.

Van de Vijver, F., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Erlbaum.

Webb, N. (1997). *Alignment of science and mathematics standards and assessments in four states*. Research Monograph Number 6: CCSSO.

Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America, 17*, 684–688.

Wright, B. D. (1999). Rasch measurement models. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 85–97). Elsevier Science Ltd.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*(1), 115–148. doi:10.1207/s15327906mbr4001_5

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS RR-12-08). ETS.

Appendix A: Spanish Pilot User Norms

Table A.1. Achievement Status Norms and Fall-to-Spring Growth Norms—Grade K

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
1	119	120	12.6	10.4
2	121	123	12.6	10.4
3	122	125	12.6	10.4
4	123	126	12.6	10.4
5	123	127	12.6	10.4
6	124	129	12.6	10.4
7	125	130	12.6	10.4
8	125	130	12.6	10.4
9	126	131	12.6	10.4
10	126	132	12.6	10.4
11	127	133	12.6	10.4
12	127	133	12.6	10.4
13	127	134	12.6	10.4
14	128	135	12.6	10.4
15	128	135	12.6	10.4
16	128	136	12.6	10.4
17	129	136	12.6	10.4
18	129	137	12.6	10.4
19	129	137	12.6	10.4
20	130	138	12.6	10.4
21	130	138	12.6	10.4
22	130	139	12.6	10.4
23	130	139	12.6	10.4
24	131	139	12.6	10.4
25	131	140	12.6	10.4
26	131	140	12.6	10.4
27	131	141	12.6	10.4
28	131	141	12.6	10.4
29	132	141	12.6	10.4
30	132	142	12.6	10.4
31	132	142	12.6	10.4
32	132	142	12.6	10.4
33	133	143	12.6	10.4
34	133	143	12.6	10.4
35	133	144	12.6	10.4
36	133	144	12.6	10.4
37	133	144	12.6	10.4
38	134	145	12.6	10.4
39	134	145	12.6	10.4
40	134	145	12.6	10.4
41	134	146	12.6	10.4
42	134	146	12.6	10.4

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
43	135	146	12.6	10.4
44	135	147	12.6	10.4
45	135	147	12.6	10.4
46	135	147	12.6	10.4
47	135	148	12.6	10.4
48	136	148	12.6	10.4
49	136	148	12.6	10.4
50	136	148	12.6	10.4
51	136	149	12.6	10.4
52	136	149	12.6	10.4
53	136	149	12.6	10.4
54	137	150	12.6	10.4
55	137	150	12.6	10.4
56	137	150	12.6	10.4
57	137	151	12.6	10.4
58	137	151	12.6	10.4
59	138	151	12.6	10.4
60	138	152	12.6	10.4
61	138	152	12.6	10.4
62	138	152	12.6	10.4
63	138	153	12.6	10.4
64	139	153	12.6	10.4
65	139	153	12.6	10.4
66	139	154	12.6	10.4
67	139	154	12.6	10.4
68	139	154	12.6	10.4
69	140	155	12.6	10.4
70	140	155	12.6	10.4
71	140	156	12.6	10.4
72	140	156	12.6	10.4
73	141	156	12.6	10.4
74	141	157	12.6	10.4
75	141	157	12.6	10.4
76	141	158	12.6	10.4
77	142	158	12.6	10.4
78	142	158	12.6	10.4
79	142	159	12.6	10.4
80	142	159	12.6	10.4
81	143	160	12.6	10.4
82	143	160	12.6	10.4
83	143	161	12.6	10.4
84	143	161	12.6	10.4
85	144	162	12.6	10.4
86	144	162	12.6	10.4
87	144	163	12.6	10.4

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
88	145	164	12.6	10.4
89	145	164	12.6	10.4
90	146	165	12.6	10.4
91	146	166	12.6	10.4
92	147	167	12.6	10.4
93	147	167	12.6	10.4
94	148	168	12.6	10.4
95	148	170	12.6	10.4
96	149	171	12.6	10.4
97	150	173	12.6	10.4
98	152	176	12.6	10.4
99	153	177	12.5	10.4

Table A.2. Achievement Status Norms and Fall-to-Spring Growth Norms—Grade 1

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
1	122	132	17.4	9.5
2	125	135	17.0	9.5
3	127	137	16.8	9.5
4	128	139	16.6	9.5
5	129	140	16.5	9.5
6	130	141	16.4	9.5
7	131	142	16.3	9.5
8	132	143	16.1	9.5
9	133	144	16.0	9.5
10	134	145	15.9	9.5
11	134	146	15.9	9.5
12	135	147	15.7	9.5
13	136	147	15.6	9.5
14	136	148	15.6	9.5
15	137	149	15.5	9.5
16	137	149	15.5	9.5
17	138	150	15.3	9.5
18	138	150	15.3	9.5
19	139	151	15.2	9.5
20	139	151	15.2	9.5
21	140	152	15.1	9.5
22	140	152	15.1	9.5
23	141	153	14.9	9.5
24	141	153	14.9	9.5
25	141	154	14.9	9.5
26	142	154	14.8	9.5
27	142	155	14.8	9.5
28	142	155	14.8	9.5

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
29	143	156	14.7	9.5
30	143	156	14.7	9.5
31	144	156	14.5	9.5
32	144	157	14.5	9.5
33	144	157	14.5	9.5
34	145	158	14.4	9.5
35	145	158	14.4	9.5
36	145	158	14.4	9.5
37	146	159	14.3	9.5
38	146	159	14.3	9.5
39	146	160	14.3	9.5
40	147	160	14.1	9.5
41	147	160	14.1	9.5
42	147	161	14.1	9.5
43	148	161	14.0	9.5
44	148	161	14.0	9.5
45	148	162	14.0	9.5
46	149	162	13.9	9.5
47	149	162	13.9	9.5
48	149	163	13.9	9.5
49	149	163	13.9	9.5
50	150	164	13.7	9.5
51	150	164	13.7	9.5
52	150	164	13.7	9.5
53	151	165	13.6	9.5
54	151	165	13.6	9.5
55	151	165	13.6	9.5
56	152	166	13.5	9.5
57	152	166	13.5	9.5
58	152	166	13.5	9.5
59	153	167	13.3	9.5
60	153	167	13.3	9.5
61	153	168	13.3	9.5
62	154	168	13.2	9.5
63	154	168	13.2	9.5
64	154	169	13.2	9.5
65	155	169	13.1	9.5
66	155	170	13.1	9.5
67	155	170	13.1	9.5
68	156	170	12.9	9.5
69	156	171	12.9	9.5
70	156	171	12.9	9.5
71	157	172	12.8	9.5
72	157	172	12.8	9.5
73	157	172	12.8	9.5

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
74	158	173	12.7	9.5
75	158	173	12.7	9.5
76	159	174	12.6	9.5
77	159	174	12.6	9.5
78	159	175	12.6	9.5
79	160	175	12.4	9.5
80	160	176	12.4	9.5
81	161	176	12.3	9.5
82	161	177	12.3	9.5
83	162	177	12.2	9.5
84	162	178	12.2	9.5
85	163	179	12.0	9.5
86	163	179	12.0	9.5
87	164	180	11.9	9.5
88	165	181	11.8	9.5
89	165	181	11.8	9.5
90	166	182	11.6	9.5
91	167	183	11.5	9.5
92	167	184	11.5	9.5
93	168	185	11.4	9.5
94	169	186	11.2	9.5
95	171	188	11.0	9.5
96	172	189	10.8	9.5
97	174	191	10.6	9.5
98	177	194	10.2	9.5
99	178	195	10.0	9.5

Table A.3. Achievement Status Norms and Fall-to-Spring Growth Norms—Grade 2

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
1	137	143	12.5	9.1
2	140	146	12.3	9.1
3	142	149	12.1	9.1
4	144	150	11.9	9.1
5	145	152	11.8	9.1
6	146	153	11.8	9.1
7	147	154	11.7	9.1
8	148	155	11.6	9.1
9	149	156	11.5	9.1
10	150	157	11.4	9.1
11	150	158	11.4	9.1
12	151	159	11.3	9.1
13	152	160	11.3	9.1
14	152	160	11.3	9.1

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
15	153	161	11.2	9.1
16	153	162	11.2	9.1
17	154	162	11.1	9.1
18	154	163	11.1	9.1
19	155	163	11.0	9.1
20	155	164	11.0	9.1
21	156	164	10.9	9.1
22	156	165	10.9	9.1
23	157	165	10.9	9.1
24	157	166	10.9	9.1
25	158	166	10.8	9.1
26	158	167	10.8	9.1
27	158	167	10.8	9.1
28	159	168	10.7	9.1
29	159	168	10.7	9.1
30	160	169	10.6	9.1
31	160	169	10.6	9.1
32	160	170	10.6	9.1
33	161	170	10.5	9.1
34	161	170	10.5	9.1
35	161	171	10.5	9.1
36	162	171	10.4	9.1
37	162	172	10.4	9.1
38	163	172	10.4	9.1
39	163	172	10.4	9.1
40	163	173	10.4	9.1
41	164	173	10.3	9.1
42	164	174	10.3	9.1
43	164	174	10.3	9.1
44	165	174	10.2	9.1
45	165	175	10.2	9.1
46	165	175	10.2	9.1
47	166	176	10.1	9.1
48	166	176	10.1	9.1
49	166	176	10.1	9.1
50	167	177	10.0	9.1
51	167	177	10.0	9.1
52	167	177	10.0	9.1
53	168	178	9.9	9.1
54	168	178	9.9	9.1
55	168	179	9.9	9.1
56	169	179	9.9	9.1
57	169	179	9.9	9.1
58	169	180	9.9	9.1
59	170	180	9.8	9.1

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
60	170	180	9.8	9.1
61	170	181	9.8	9.1
62	171	181	9.7	9.1
63	171	182	9.7	9.1
64	171	182	9.7	9.1
65	172	182	9.6	9.1
66	172	183	9.6	9.1
67	172	183	9.6	9.1
68	173	184	9.5	9.1
69	173	184	9.5	9.1
70	174	185	9.4	9.1
71	174	185	9.4	9.1
72	174	185	9.4	9.1
73	175	186	9.4	9.1
74	175	186	9.4	9.1
75	176	187	9.3	9.1
76	176	187	9.3	9.1
77	176	188	9.3	9.1
78	177	188	9.2	9.1
79	177	189	9.2	9.1
80	178	189	9.1	9.1
81	178	190	9.1	9.1
82	179	191	9.0	9.1
83	179	191	9.0	9.1
84	180	192	8.9	9.1
85	180	192	8.9	9.1
86	181	193	8.9	9.1
87	181	194	8.9	9.1
88	182	194	8.8	9.1
89	183	195	8.7	9.1
90	184	196	8.6	9.1
91	184	197	8.6	9.1
92	185	198	8.5	9.1
93	186	199	8.4	9.1
94	187	200	8.4	9.1
95	189	202	8.2	9.1
96	190	204	8.1	9.1
97	192	206	7.9	9.1
98	195	209	7.7	9.1
99	196	210	7.6	9.1

Table A.4. Achievement Status Norms and Fall-to-Spring Growth Norms—Grade 3

	Status Norms		F-S Cond. Growth Norms	
Percentile	**Fall**	**Spring**	**E(S-F|F)**	**SD(S-F|F)**
1	146	151	11.0	9.4
2	149	155	10.8	9.4
3	152	157	10.6	9.4
4	153	159	10.6	9.4
5	155	161	10.5	9.4
6	156	162	10.4	9.4
7	157	163	10.3	9.4
8	158	164	10.3	9.4
9	159	165	10.2	9.4
10	160	166	10.1	9.4
11	161	167	10.1	9.4
12	161	168	10.1	9.4
13	162	169	10.0	9.4
14	163	170	10.0	9.4
15	163	170	10.0	9.4
16	164	171	9.9	9.4
17	165	172	9.8	9.4
18	165	172	9.8	9.4
19	166	173	9.8	9.4
20	166	174	9.8	9.4
21	167	174	9.7	9.4
22	167	175	9.7	9.4
23	168	175	9.7	9.4
24	168	176	9.7	9.4
25	169	176	9.6	9.4
26	169	177	9.6	9.4
27	170	177	9.5	9.4
28	170	178	9.5	9.4
29	170	178	9.5	9.4
30	171	179	9.5	9.4
31	171	179	9.5	9.4
32	172	180	9.4	9.4
33	172	180	9.4	9.4
34	173	181	9.3	9.4
35	173	181	9.3	9.4
36	173	182	9.3	9.4
37	174	182	9.3	9.4
38	174	182	9.3	9.4
39	174	183	9.3	9.4
40	175	183	9.2	9.4
41	175	184	9.2	9.4
42	176	184	9.2	9.4
43	176	185	9.2	9.4

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
44	176	185	9.2	9.4
45	177	185	9.1	9.4
46	177	186	9.1	9.4
47	177	186	9.1	9.4
48	178	187	9.0	9.4
49	178	187	9.0	9.4
50	179	188	9.0	9.4
51	179	188	9.0	9.4
52	179	188	9.0	9.4
53	180	189	8.9	9.4
54	180	189	8.9	9.4
55	180	190	8.9	9.4
56	181	190	8.8	9.4
57	181	190	8.8	9.4
58	182	191	8.8	9.4
59	182	191	8.8	9.4
60	182	192	8.8	9.4
61	183	192	8.7	9.4
62	183	193	8.7	9.4
63	183	193	8.7	9.4
64	184	194	8.7	9.4
65	184	194	8.7	9.4
66	185	194	8.6	9.4
67	185	195	8.6	9.4
68	185	195	8.6	9.4
69	186	196	8.5	9.4
70	186	196	8.5	9.4
71	187	197	8.5	9.4
72	187	197	8.5	9.4
73	188	198	8.4	9.4
74	188	198	8.4	9.4
75	188	199	8.4	9.4
76	189	199	8.3	9.4
77	189	200	8.3	9.4
78	190	200	8.3	9.4
79	190	201	8.3	9.4
80	191	202	8.2	9.4
81	191	202	8.2	9.4
82	192	203	8.2	9.4
83	193	203	8.1	9.4
84	193	204	8.1	9.4
85	194	205	8.0	9.4
86	194	206	8.0	9.4
87	195	206	8.0	9.4
88	196	207	7.9	9.4

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
89	197	208	7.8	9.4
90	197	209	7.8	9.4
91	198	210	7.8	9.4
92	199	211	7.7	9.4
93	200	212	7.7	9.4
94	201	214	7.6	9.4
95	203	215	7.5	9.4
96	205	217	7.3	9.4
97	207	220	7.2	9.4
98	210	223	7.0	9.4
99	211	224	7.0	9.4

Table A.5. Achievement Status Norms and Fall-to-Spring Growth Norms—Grade 4

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
1	154	158	10.1	9.5
2	157	161	9.8	9.5
3	159	164	9.7	9.5
4	161	166	9.5	9.5
5	163	167	9.3	9.5
6	164	169	9.2	9.5
7	165	170	9.1	9.5
8	166	171	9.1	9.5
9	167	172	9.0	9.5
10	168	173	8.9	9.5
11	169	174	8.8	9.5
12	170	175	8.7	9.5
13	171	176	8.6	9.5
14	171	177	8.6	9.5
15	172	178	8.5	9.5
16	173	178	8.5	9.5
17	173	179	8.5	9.5
18	174	180	8.4	9.5
19	174	180	8.4	9.5
20	175	181	8.3	9.5
21	176	181	8.2	9.5
22	176	182	8.2	9.5
23	177	183	8.1	9.5
24	177	183	8.1	9.5
25	178	184	8.0	9.5
26	178	184	8.0	9.5
27	179	185	8.0	9.5
28	179	185	8.0	9.5
29	179	186	8.0	9.5

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
30	180	186	7.9	9.5
31	180	187	7.9	9.5
32	181	187	7.8	9.5
33	181	188	7.8	9.5
34	182	188	7.7	9.5
35	182	189	7.7	9.5
36	183	189	7.6	9.5
37	183	190	7.6	9.5
38	183	190	7.6	9.5
39	184	190	7.5	9.5
40	184	191	7.5	9.5
41	185	191	7.4	9.5
42	185	192	7.4	9.5
43	185	192	7.4	9.5
44	186	193	7.4	9.5
45	186	193	7.4	9.5
46	187	194	7.3	9.5
47	187	194	7.3	9.5
48	187	194	7.3	9.5
49	188	195	7.2	9.5
50	188	195	7.2	9.5
51	188	196	7.2	9.5
52	189	196	7.1	9.5
53	189	197	7.1	9.5
54	190	197	7.0	9.5
55	190	197	7.0	9.5
56	190	198	7.0	9.5
57	191	198	6.9	9.5
58	191	199	6.9	9.5
59	192	199	6.8	9.5
60	192	200	6.8	9.5
61	192	200	6.8	9.5
62	193	201	6.8	9.5
63	193	201	6.8	9.5
64	194	201	6.7	9.5
65	194	202	6.7	9.5
66	195	202	6.6	9.5
67	195	203	6.6	9.5
68	195	203	6.6	9.5
69	196	204	6.5	9.5
70	196	204	6.5	9.5
71	197	205	6.4	9.5
72	197	205	6.4	9.5
73	198	206	6.3	9.5
74	198	206	6.3	9.5

	Status Norms		F-S Cond. Growth Norms	
Percentile	**Fall**	**Spring**	**E(S-F|F)**	**SD(S-F|F)**
75	199	207	6.3	9.5
76	199	207	6.3	9.5
77	200	208	6.2	9.5
78	200	208	6.2	9.5
79	201	209	6.1	9.5
80	201	210	6.1	9.5
81	202	210	6.0	9.5
82	202	211	6.0	9.5
83	203	212	5.9	9.5
84	204	212	5.8	9.5
85	204	213	5.8	9.5
86	205	214	5.7	9.5
87	206	215	5.7	9.5
88	206	215	5.7	9.5
89	207	216	5.6	9.5
90	208	217	5.5	9.5
91	209	218	5.4	9.5
92	210	219	5.3	9.5
93	211	221	5.2	9.5
94	212	222	5.1	9.5
95	214	224	5.0	9.5
96	216	226	4.8	9.5
97	218	228	4.6	9.5
98	221	232	4.4	9.5
99	222	233	4.3	9.5

Table A.6. Achievement Status Norms and Fall-to-Spring Growth Norms—Grade 5

	Status Norms		F-S Cond. Growth Norms	
Percentile	**Fall**	**Spring**	**E(S-F|F)**	**SD(S-F|F)**
1	156	161	10.2	8.9
2	160	165	9.7	8.9
3	162	167	9.5	8.9
4	164	169	9.2	8.9
5	166	171	9.0	8.9
6	167	172	8.9	8.9
7	169	174	8.6	8.9
8	170	175	8.5	8.9
9	171	176	8.4	8.9
10	172	177	8.3	8.9
11	173	178	8.2	8.9
12	174	179	8.1	8.9
13	174	180	8.1	8.9
14	175	180	7.9	8.9
15	176	181	7.8	8.9

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
16	177	182	7.7	8.9
17	177	183	7.7	8.9
18	178	183	7.6	8.9
19	179	184	7.5	8.9
20	179	185	7.5	8.9
21	180	185	7.3	8.9
22	180	186	7.3	8.9
23	181	186	7.2	8.9
24	182	187	7.1	8.9
25	182	187	7.1	8.9
26	183	188	7.0	8.9
27	183	189	7.0	8.9
28	184	189	6.9	8.9
29	184	190	6.9	8.9
30	185	190	6.7	8.9
31	185	191	6.7	8.9
32	186	191	6.6	8.9
33	186	192	6.6	8.9
34	186	192	6.6	8.9
35	187	192	6.5	8.9
36	187	193	6.5	8.9
37	188	193	6.4	8.9
38	188	194	6.4	8.9
39	189	194	6.3	8.9
40	189	195	6.3	8.9
41	190	195	6.1	8.9
42	190	196	6.1	8.9
43	190	196	6.1	8.9
44	191	197	6.0	8.9
45	191	197	6.0	8.9
46	192	197	5.9	8.9
47	192	198	5.9	8.9
48	193	198	5.8	8.9
49	193	199	5.8	8.9
50	193	199	5.8	8.9
51	194	200	5.7	8.9
52	194	200	5.7	8.9
53	195	200	5.6	8.9
54	195	201	5.6	8.9
55	196	201	5.4	8.9
56	196	202	5.4	8.9
57	196	202	5.4	8.9
58	197	203	5.3	8.9
59	197	203	5.3	8.9
60	198	204	5.2	8.9

	Status Norms		F-S Cond. Growth Norms	
Percentile	Fall	Spring	E(S-F|F)	SD(S-F|F)
61	198	204	5.2	8.9
62	199	204	5.1	8.9
63	199	205	5.1	8.9
64	199	205	5.1	8.9
65	200	206	5.0	8.9
66	200	206	5.0	8.9
67	201	207	4.8	8.9
68	201	207	4.8	8.9
69	202	208	4.7	8.9
70	202	208	4.7	8.9
71	203	209	4.6	8.9
72	203	209	4.6	8.9
73	204	210	4.5	8.9
74	204	210	4.5	8.9
75	205	211	4.4	8.9
76	205	211	4.4	8.9
77	206	212	4.2	8.9
78	206	212	4.2	8.9
79	207	213	4.1	8.9
80	208	214	4.0	8.9
81	208	214	4.0	8.9
82	209	215	3.9	8.9
83	209	216	3.9	8.9
84	210	216	3.8	8.9
85	211	217	3.6	8.9
86	212	218	3.5	8.9
87	212	219	3.5	8.9
88	213	219	3.4	8.9
89	214	220	3.3	8.9
90	215	221	3.2	8.9
91	216	222	3.1	8.9
92	217	224	2.9	8.9
93	218	225	2.8	8.9
94	220	226	2.6	8.9
95	221	228	2.5	8.9
96	223	230	2.2	8.9
97	226	233	2.0	8.9
98	229	236	2.0	8.9
99	230	237	2.0	8.9

Table A.7. Spring Status Percentiles—Grades 6–8

	Spring Status Percentiles		
Percentile	Grade 6	Grade 7	Grade 8
1	173	175	174
2	176	179	178
3	178	181	180
4	180	183	182
5	181	184	184
6	182	186	186
7	183	187	187
8	184	188	188
9	185	189	189
10	186	190	190
11	187	190	191
12	187	191	192
13	188	192	193
14	189	193	193
15	189	193	194
16	190	194	195
17	191	195	196
18	191	195	196
19	192	196	197
20	192	196	197
21	193	197	198
22	193	198	199
23	193	198	199
24	194	199	200
25	194	199	200
26	195	200	201
27	195	200	201
28	196	200	202
29	196	201	202
30	196	201	203
31	197	202	203
32	197	202	204
33	198	203	204
34	198	203	205
35	198	204	205
36	199	204	206
37	199	204	206
38	199	205	207
39	200	205	207
40	200	206	208
41	201	206	208
42	201	206	208
43	201	207	209
44	202	207	209

	Spring Status Percentiles		
Percentile	Grade 6	Grade 7	Grade 8
45	202	208	210
46	202	208	210
47	203	208	211
48	203	209	211
49	203	209	211
50	204	209	212
51	204	210	212
52	204	210	213
53	205	211	213
54	205	211	214
55	205	211	214
56	206	212	214
57	206	212	215
58	207	213	215
59	207	213	216
60	207	213	216
61	208	214	217
62	208	214	217
63	208	215	218
64	209	215	218
65	209	215	218
66	209	216	219
67	210	216	219
68	210	217	220
69	211	217	220
70	211	218	221
71	211	218	221
72	212	219	222
73	212	219	222
74	213	219	223
75	213	220	223
76	213	220	224
77	214	221	224
78	214	221	225
79	215	222	226
80	215	223	226
81	216	223	227
82	216	224	227
83	217	224	228
84	217	225	229
85	218	226	230
86	219	226	230
87	219	227	231
88	220	228	232
89	221	229	233

	Spring Status Percentiles		
Percentile	Grade 6	Grade 7	Grade 8
90	221	229	234
91	222	230	235
92	223	231	236
93	224	232	237
94	225	234	239
95	227	235	240
96	228	237	242
97	230	239	245
98	233	243	248
99	234	244	249