MAP Growth Technical Report
Spanish Addendum for 2024–2025

April 2, 2026

MAP Growth Technical Report Spanish Addendum for 2024–2025

Table of Contents

List of Tables

List of Figures

1. Introduction

Spanish-language versions of the MAP Growth assessments are available for Math and Reading. These Spanish tests parallel the English versions of MAP Growth in their design, implementation, scoring, and reporting. This addendum to the *MAP Growth Technical Report for 2024–2025* (NWEA, 2026) focuses on aspects of the Spanish assessments that differ from their English counterparts. Unless otherwise stated in this report, all aspects of MAP Growth Spanish assessments are the same as the English assessments, whose features are fully described in the main MAP Growth technical report. This addendum also presents psychometric evidence specific to the Spanish Math and Spanish Reading assessments, as students taking the Spanish versions of MAP Growth represent a different population than those taking the English versions.

There is no need to duplicate the information in this addendum for those sections where the content in the main report is identical or encompasses Spanish Math and Spanish Reading alongside the English-language assessments. Where applicable, chapter, section, table, and figure titles in this addendum match those in the main report to facilitate comparison between the two documents.

The Spanish and English versions of the MAP Growth Math assessments are very similar and exist on the same scale for two main reasons: First, the mathematics construct and content domains are the same in either language. Second, all Spanish Math items are transadaptations of their counterpart English Math items. By design, the language of the items is the only difference between the two Math assessments.

The Spanish and English Reading assessments inherently have more differences because of the nature of reading in each language. The MAP Growth Spanish Reading assessment consists of Spanish-specific content and uses a Spanish-specific reading scale and Spanish-specific norms. While some items on the Spanish Reading assessment are transadapted from the English Reading pool, the majority of items have been developed in Spanish by Spanish language arts content specialists. Linguistic and content differences drove the decision to have separate scales for MAP Growth Reading in Spanish and English. For example, English orthography is extremely complex and inconsistent as compared with the transparent orthography of Spanish. Linguistic differences also suggest that Spanish may bring shifts in relative difficulty across phonics, phonological awareness, spelling, and achievement of accurate decoding as compared with English. The need for separate Reading assessments for each language is also borne out both by research on reading development and by reviews of English-specific and Spanish-specific standards.

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

As shown in Table 2.1 and Table 2.2, NWEA offers Spanish-language versions of MAP Growth Math and Reading tests. These tests are designed for students who speak Spanish as their first language or receive instruction in Spanish. Within the Spanish-language assessments, all item content—including passages in the Spanish Reading assessments, audio on the K–2 tests, captions or labels on graphics, and item directions—is presented to the student in Spanish. These Spanish Math and Spanish Reading assessments are computer adaptive tests that can be administered up to four times per school year—in fall, winter, spring, and with an optional fourth administration in summer.

While the MAP Growth Spanish Math assessments use the same scale and norms as the English Math assessments, MAP Growth Spanish Reading is on its own scale and has independent norms. Reporting features for both Spanish Math and Spanish Reading assessments are consistent with the English assessments. Educators can receive scale score data from both English-language MAP Growth Reading and MAP Growth Spanish Reading if students take both assessments, empowering them to make informed decisions to support their students' learning in both languages. For Math, if a student takes a Spanish test and an English test in the same test administration window, the test with the lower standard error will be the one reported and used for growth projections because both tests provide scores on the same scale. The design of the MAP Growth Spanish Math and Spanish Reading assessments was guided by the same underlying principles as their English-version counterparts, including Universal Design for Learning (UDL) principles (Thompson et al., 2002).

2.1. Spanish Math Test Design

MAP Growth Spanish Math tests support assessment of mathematical achievement and growth without the barrier of English-language fluency across grades K–12. These tests use the same blueprints and designs as the English-language MAP Growth Math tests. These shared blueprints and their associated tests are grade-banded: MAP Growth Spanish Math tests are available for grades K–2, 2–5, and 6+. The blueprints used by both English and Spanish Math tests represent the content domain defined by standards in mathematics and delineate the cross-grade instructional areas that represent the top level of standards, such as the domain level (e.g., Algebra for Math). Instructional areas for Spanish Math are the same for every grade in the grade band, but standards are specific to each grade. Grade-specific standards are mapped to each instructional area.

The item pool for each Spanish Math assessment is created by choosing items from the Spanish Math item bank and aligning each selected item to a mathematics standard and instructional area. The grade level of an item is assigned according to the grade level of the standard. The items in the Spanish Math item pools are aligned to the same standards as English-language Math items, and alignment decisions are made using a consistent alignment philosophy and process between both English Math and Spanish Math. Each Spanish Math item has the same alignment as its English Math counterpart item.

As shown in Table 2.1, the blueprints shared between English and Spanish Math assessments indicate the target number of items per instructional area. The number of items in each instructional area is chosen in proportion to that domain's emphasis in the standards or content framework to which the test is aligned, with a maximum of 43 items for each test event. In addition, the shared Math blueprints define the minimum number of items associated with each

category in the Aspects of Rigor (AOR) framework, which characterizes mathematical cognitive complexity developed by Achieve (2019).

Table 2.1. Example Spanish Math Blueprint

Test Name	Content Feature	Instructional Area Score	Number of Items
Spanish Math 2–5	Geometry	Yes	6
	Measurement and Data	Yes	11
	Number and Operations	Yes	13
	Operations and Algebraic Thinking	Yes	10
	AOR: Procedural	No	At least 13
	AOR: Conceptual	No	At least 16
	AOR: Application	No	At least 6

Note. AOR = Aspects of Rigor

2.2. Spanish Reading Test Design

MAP Growth Spanish Reading assessments measure students' current Spanish reading achievement and track longitudinal growth of reading achievement in grades K–8. While the structures of the Spanish and English Reading tests are closely connected, the two assessments have separate scales and norms due to the inherent differences in construct between English and Spanish literacy, including the rate of literacy development. When investigating the degree of construct equivalence between English literacy and Spanish literacy, the research literature points to significant differences within the learning-to-read space (Jiban, 2017). These largely derive from the extremely complex and inconsistent orthography in English as compared with the 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. However, once independent reading comprehension becomes the focus of assessment, the differences between the two languages and the relative constructs are less distinguishable. These broad findings are borne out both by research on reading development and by reviews of English-specific and Spanish-specific standards.

Each MAP Growth Spanish Reading test is defined by content area and grade band. MAP Growth Spanish 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: reading foundational skills, reading comprehension, vocabulary, and language and writing standards. In contrast, students who take the 2–5 and 6–8 tests have progressed to independent reading, and the assessment targets reading comprehension, understanding of genres and text, and vocabulary. The split between the 2–5 and 6–8 tests helps ensure that students see content appropriate to their age and performance 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.

The Spanish Reading blueprints represent the content domain defined by academic content standards (either English language arts or Spanish language arts) and delineate the cross-grade instructional areas that represent the top level of standards, such as the domain level (e.g., Reading Informational for reading). Instructional areas for Spanish Reading are the same

for every grade in the grade band, but standards are specific to each grade. The instructional areas and sub-areas are written in either English or Spanish, matching the language of the standards the test is aligned to. Grade-specific standards are mapped to each instructional area. The instructional areas and sub-areas, as well as the standard mapping of Spanish Reading assessments, generally align with their English counterparts. Variations may occur where standards themselves differ, such as when states adopt Spanish language arts standards that include specifics about Spanish orthography, morphology, and language conventions. When developing Spanish tests aligned to standards that are not explicitly designed to incorporate Spanish literacy development, items that represent Spanish-specific constructs are included in the assessments by aligning items to the higher level ("parent") standards or aligning to standards that specify "standard English."

The item pool for each Spanish Reading assessment is created by choosing items from the Spanish Reading item bank and aligning each selected item to an English language arts or Spanish language arts standard and instructional area. The grade level of an item is assigned according to the grade level of the standard. The items in the Spanish Reading item pools are aligned independently from English Reading items, although a similar and systematic alignment philosophy and process guides decisions for both scales.

As shown in Table 2.2, the blueprints for Spanish Reading assessments indicate the number of items per instructional area. Items in Spanish Reading assessments are evenly distributed across instructional areas, with a maximum of 43 items for each test event. An additional content feature driving item selection for the 2–5 and 6–8 Spanish Reading tests is the inclusion of an item set that contains an extended reading passage and 3 questions about the passage. This item set currently includes only field test items, while the corresponding item sets in English-language Reading 2–5 and 6+ tests include both operational and field test items. The Spanish Reading test design will be updated to include both operational and field test item sets once a sufficient pool of set-based items is calibrated.

Table 2.2. Example Spanish Reading Test Blueprint

Test Name	Content Feature	Instructional Area Score	Number of Items
Spanish Reading 2–5	Texto informativo	Yes	13
	Texto literario	Yes	13
	Vocabulario	Yes	13
	Item Set with Reading Passage	No	Up to 1

3. Item Development Follows a Rigorous Process

3.1. Transadaptation Process

The Spanish Math item bank is composed entirely of items that have been transadapted using a careful and highly structured process that starts with calibrated English Math items. The Spanish Reading bank was initially composed of similarly transadapted items, but the majority of that bank now consists of items developed in Spanish rather than having been transadapted.

Transadaptation for both Spanish Math and, when used previously, Spanish Reading follows the same process. Central to this approach is the understanding that the Spanish Math and Spanish Reading items are not only *translations* but *adaptations* of English items. Original English content is adapted to be culturally and linguistically appropriate in Spanish. Items selected for transadaptation have strong content and alignment and 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 consider the following:

- Language variation related to regionalisms and dialects
- Differences 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 the construct to target the skill of the original item

The International Test Commission guidelines inform the workflow and steps NWEA follows in transadapting items to help achieve similarity and parallelism (International Test Commission, 2017). The English and Spanish Reading tests assess similar concepts in their respective languages, while the English and Spanish Math tests assess the same concepts in both languages. The transadaptation process aims to

- define the target audience and age group,
- define the 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, and terminology, and
- achieve test equivalence.

To conduct item development via transadaptation, NWEA content specialists select a possible pool of items to be transadapted. A glossary of terms is created for consistency, and extraction of both item and image text is used to translate and transadapt new terms and add to the glossary. Each selected English item is analyzed to determine whether to: translate the item where possible, as with simple computation items; transadapt the item where necessary, as when context and linguistic complexity is present; or remove the item from the development project as not a viable candidate. After the translation and/or transadaptation of item content and images, items undergo reviews for linguistic editing, content editing, and proofreading.

- Linguistic editing:
    - Edit for grammar, syntax, style, and flow

    - Ensure that concept and meaning are 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's 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 validate the transadaptations and resolve any issues. All items undergo review and revision, if needed, before a final content review to ensure that the new Spanish-language item preserves the quality of the original content, requires the same knowledge and abilities needed to answer the original item, represents the standards and probable instruction of the target population, and is free of errors, flaws, or multiple keys. Transadapted items are entered into the NWEA content management system, and a final quality check is conducted to make sure the English and Spanish items display alike and that the metadata is accurate to the Spanish-language item.

3.2. Spanish-Language Development Process

While Spanish Math items continue to be developed via transadaptation, Spanish Reading items are now originally created in Spanish following the same rigorous and multi-stage development process described in the main MAP Growth technical report, with some small modifications. For example, where determined as appropriate by Spanish content specialists, existing English Reading standards-based item specifications may be leveraged when the English language arts and Spanish language arts standards match and no additional language concerns are present. In addition, the same detailed metadata is applied to Spanish Reading items as for other items, with adjustments for substituting readability measures that are relevant to Spanish-language text complexity. Using consistent development and review steps focused on content validity and quality across all scales, including Spanish Reading, ensures high-quality items are published and exposed to students.

3.3. Passage Development

Text excerpts are used with MAP Growth Spanish Reading items. Some are short passages attached to standalone items, whereas others are extended texts that can support a set of 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 text-based items that require close reading of the extended text and are therefore included in the item bank to address concepts and state standards that require complex texts.

All texts used with Spanish Reading items are either transadapted versions of commissioned passages from the English Reading bank, fully original Spanish-language commissioned informational and literary passages, or copyrighted works used with permission. Using fully original Spanish commissioned and copyrighted works supports the cultural relevance and accessibility of the texts for the target population. For commissioned informational text passages, writers provide source documentation to support fact-checking, and all passages undergo permissions and copyright review.

3.4. Field Testing, Calibration, and Psychometric Review

Spanish Reading items are field tested and calibrated using the same algorithms and processes as described in the main MAP Growth technical report for other scales. Spanish Math items, on the other hand, are not field tested because they are versioned from already field tested and calibrated English Math items and adopt the same RIT value as their English Math counterparts. The tightly controlled transadaptation process for creating Spanish Math items and the placement of the Spanish Math items on the same scale as the source English Math items allows for the use of English-calibrated item parameters to be used on the Spanish Math assessment. Wang and Li (2020) showed that the English- and Spanish-calibrated item parameters were similar and did not require separate calibrations for each language.

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

4.1. Operational Item Statistics

4.1.1. Classical Item Difficulty and Item Discrimination

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

Across all terms during the 2024–2025 test administration, Spanish Math items demonstrated a well-balanced difficulty distribution, with mean *p* values increasing from fall (0.48) to winter (0.51) and spring (0.52), indicating a modest increase in the proportion of correct responses over time and suggesting that items became slightly easier for the tested population, as shown in Table 4.1). Most items fall in the 0.4–0.7 *p*-value range, suggesting appropriate targeting for the students assessed. Correspondingly, Table 4.2 shows that Math items demonstrated strong and stable discrimination, with mean point-biserial correlations remaining constant at 0.32 across terms and the vast majority of items clustered in the 0.2–0.4 range, along with a smaller portion in the 0.4–0.6 range. The near absence of negative or near-zero discrimination values indicates that Math items consistently differentiated between lower- and higher-performing students while maintaining appropriate difficulty.

Spanish Reading items also showed an increase in mean *p* values from fall (0.44) to winter (0.46) and spring (0.48), indicating a gradual increase in the proportion of correct responses over time and suggesting somewhat easier item performance by spring, as shown in Table 4.1. The distribution is somewhat more concentrated than in Math, with most Reading items falling in the 0.3–0.6 *p*-value range, especially in the 0.4–0.5 band in fall and winter and shifting somewhat toward the 0.5–0.6 band in spring. This pattern suggests improved alignment between item difficulty and student ability over time. Table 4.2 shows that Reading also exhibited consistently strong discrimination, with mean point-biserial correlations remaining stable at 0.30 across all terms and most items concentrated in the 0.2–0.4 range, with a smaller proportion in the 0.0–0.2 and 0.4–0.6 ranges. The absence of negatively discriminating items supports that Reading items reliably measure the intended construct and effectively distinguish among performance levels.

Table 4.1. Summary of Item *P* Values

Subject	Term	N Items	Mean	SD	*P*-Value Range									
					[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
Math	Fall	1,893	0.48	0.12	0.05	0.74	6.55	16.64	31.33	30.38	10.94	2.96	0.42	0
Math	Winter	2,086	0.51	0.13	0.05	0.72	5.18	13.37	23.87	31.11	19.46	5.37	0.86	0
Math	Spring	2,257	0.52	0.12	0.04	0.44	4.30	13.25	25.56	30.84	19.72	5.01	0.84	0
Reading	Fall	1,608	0.44	0.07	0	0	1.80	26.74	55.04	14.86	1.24	0.25	0.06	0
Reading	Winter	1,775	0.46	0.07	0	0.06	1.01	16.90	53.46	25.01	2.99	0.45	0.11	0
Reading	Spring	1,906	0.48	0.08	0	0	0.94	12.49	48.48	31.16	6.24	0.63	0.05	0

Table 4.2. Summary of Item Point-Biserial Correlations

Subject	Term	N Items	Mean	SD	Point Biserial Range									
					[-1.0, -0.8]	(-0.8, -0.6]	(-0.6, -0.4]	(-0.4, -0.2]	(-0.2, 0.0]	(0.0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1.0]
Math	Fall	1,893	0.32	0.07	0	0	0	0	0.05	6.92	82.09	10.94	0	0
Math	Winter	2,086	0.32	0.07	0	0	0	0	0	5.75	82.65	11.60	0	0
Math	Spring	2,257	0.32	0.07	0	0	0	0	0.04	5.36	85.82	8.77	0	0
Reading	Fall	1,608	0.30	0.06	0	0	0	0	0	6.53	88.87	4.60	0	0
Reading	Winter	1,775	0.30	0.06	0	0	0	0	0	5.30	89.07	5.63	0	0
Reading	Spring	1,906	0.30	0.06	0	0	0	0	0	4.83	90.87	4.30	0	0

4.2. Item Fit

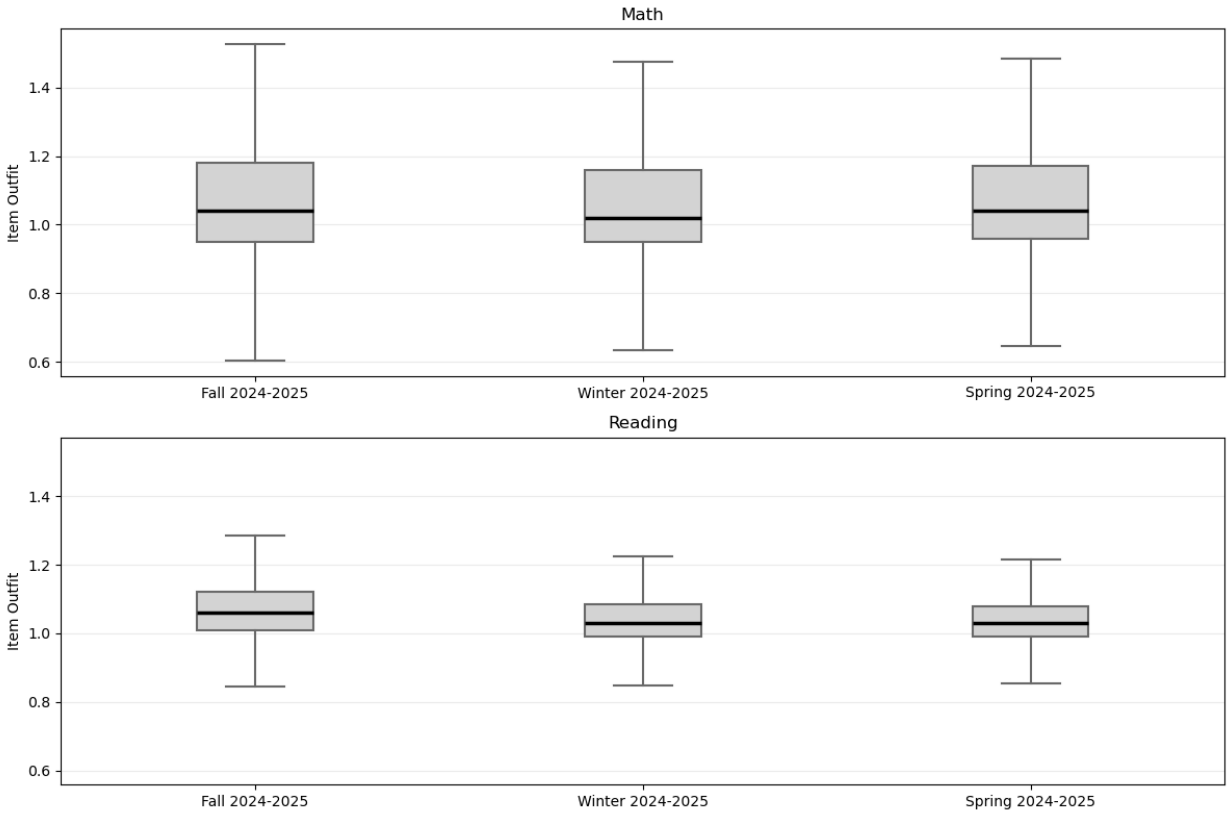
Item OUTFIT is a type of fit statistic calculated on items administered to students. The expected value of the OUTFIT statistic is 1. Values close to 1 are considered good-fitting items, with acceptable values ranging between 0.5 and 1.5; misfitting items have values outside this range.

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

4.2.1. Cross-Sectional Item Fit

The complete distribution of OUTFIT statistics for Spanish Math and Spanish Reading across fall, winter, and spring 2024–2025 is illustrated in Figure 4.1. Across terms, the boxplots show that OUTFIT values for both subjects are generally centered close to 1.0, indicating good item fit overall. Math displays slightly greater variability than Reading, with a wider spread and upper values approaching the acceptable threshold, whereas Reading shows a tighter and more consistent distribution across terms. Overall, the plots suggest that most items in both subjects fall within the acceptable fit range and that item fit remains stable over time.

Figure 4.1. Item OUTFIT Statistics by Subject and Term



4.2.2. Longitudinal Item Fit

Spanish MAP Growth items were evaluated for fit over multiple terms (i.e., fall, winter, spring 2024–2025). Table 4.3 shows the results of the item fit analyses. The vast majority of items, 91% for Math and 98% for Reading, show no misfit during any term. A small percentage shows misfit in one term (about 7% of Math items and about 1% of Reading items). Very few items show misfit in two or more terms; items that show misfit in two or more terms are likely in need of review and recalibration.

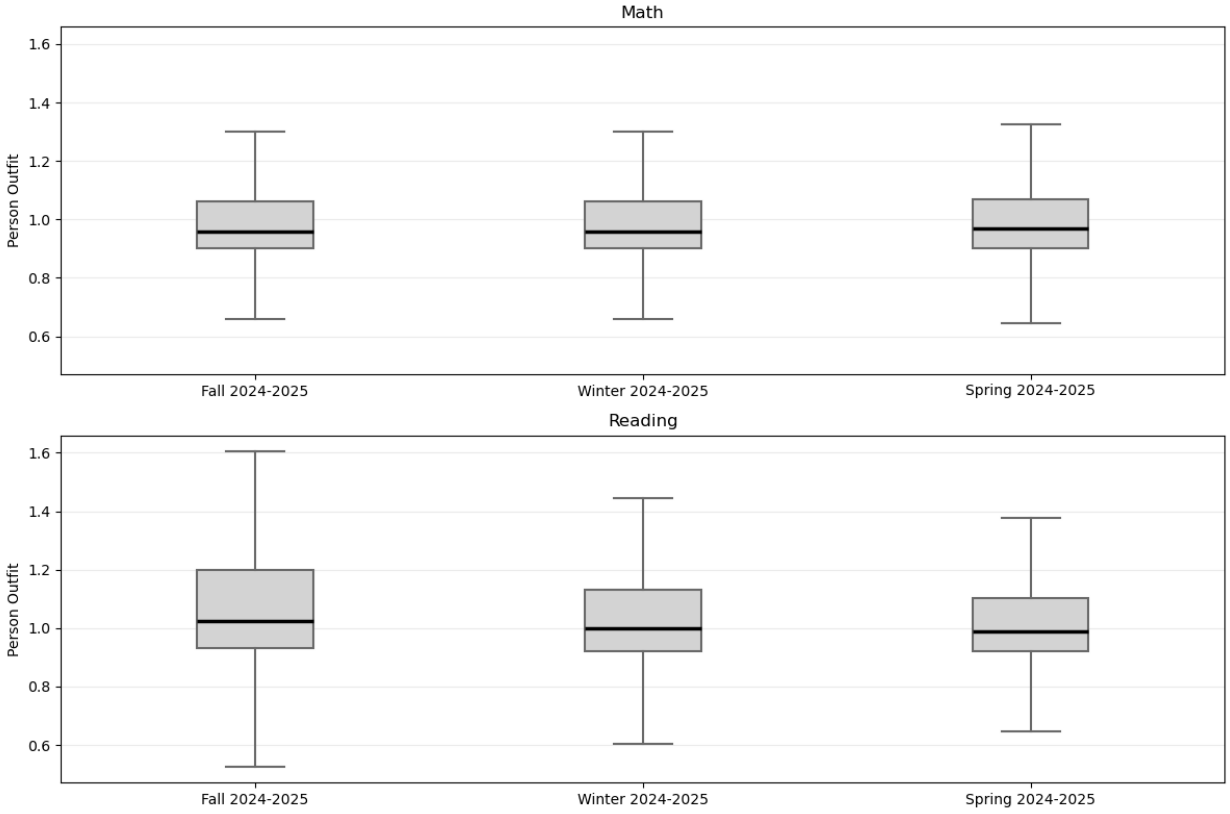
Table 4.3. Percentages of Items Misfitting Across Terms

Subject	Misfitting Terms	Number of Items	Percentage
Math	0	3,453	91.25
Math	1	253	6.69
Math	2	55	1.45
Math	3	23	0.61
Reading	0	2,419	97.70
Reading	1	31	1.25
Reading	2	16	0.65
Reading	3	10	0.40

4.3. Examinee Fit

The distribution of person OUTFIT statistics for Spanish Math and Spanish Reading across fall, winter, and spring 2024–2025 is illustrated in Figure 4.2. Across terms, the boxplots show that person OUTFIT values for both subjects centered close to 1.0, indicating good and stable examinee fit to the Rasch model. Math shows a relatively consistent distribution across administrations, with moderate spread and most values well within the acceptable range. Reading displays slightly greater variability, particularly in fall, with upper values extending closer to the threshold, but the distributions remain centered near 1.0 in all terms. Overall, the plots suggest that examinees in both subjects fit the model well and that person fit remains stable over time.

Figure 4.2. Examinee OUTFIT Statistics by Subject and Term



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

Spanish student RIT scores for 2024–2025 increased on average from kindergarten through high school, as shown in Table 4.4. Average Math scores range from a low of 158 to a high of 218. Average Reading scores range from a low of 151 to a high of 206. The average Reading score in high school is lower than the average Reading score for grade 8. The variability of scores in both subjects tends to increase by grade, with the greatest variability appearing in high school.

Item RIT values also increased on average across grades in both Math and Reading. More importantly, the item difficulty values are similar to the average student RIT scores, which indicates that items given to students were of a suitable difficulty level. The variability of item difficulty values also increased slightly by grade and had more variability in high school.

Table 4.4. Comparison of Examinee RIT Scores to Item Difficulties

Subject	Grade	Students		Items	
		Mean	SD	Mean	SD
Math	K	158.02	13.14	156.93	14.41
Math	1	173.76	14.44	172.58	15.56
Math	2	184.60	14.60	184.21	15.42
Math	3	195.43	15.94	195.15	17.01
Math	4	201.56	17.48	201.06	18.49
Math	5	204.21	18.32	203.87	19.20

Subject	Grade	Students		Items	
		Mean	SD	Mean	SD
Math	6	203.69	16.58	204.87	16.87
Math	7	206.17	18.37	207.15	18.39
Math	8	209.78	20.06	210.64	19.76
Math	HS	218.52	25.19	218.58	23.77
Reading	K	151.59	13.63	151.77	13.66
Reading	1	164.81	14.33	164.34	15.18
Reading	2	179.45	14.93	181.12	15.28
Reading	3	189.48	15.36	190.91	15.21
Reading	4	195.27	16.08	196.13	16.07
Reading	5	199.53	16.65	200.05	16.50
Reading	6	201.11	15.24	202.24	14.84
Reading	7	204.07	15.93	204.99	15.55
Reading	8	206.84	16.37	207.56	15.83
Reading	HS	201.52	18.27	203.60	16.80

Note. SD = standard deviation

4.5. Conditional Standard Error of Measurement

In item response theory, the amount of information an item provides about a student's ability is described by the item information function, where larger values indicate more-informative items. In the Rasch model, item information is defined as $I_{ij}(\theta) = P_{ij}(\theta)[1 - P_{ij}(\theta)]$. The test information function is the sum of item information across all items on a student's test, and higher test information indicates greater precision and reliability in the student's score estimate. The conditional standard error of measurement (CSEM) is defined as $CSEM_i(\theta) = 1/\sqrt{I_i(\theta)}$, making it inversely related to test information. Thus, as test information increases, CSEM decreases, meaning student scores are estimated more precisely.

A MAP Growth test event is designed to target a CSEM of about 3.3 RIT points, with slight variation by subject. This target is intended to be consistent across students, so observed CSEM values should remain fairly similar regardless of ability. As shown in Figure 4.3, average CSEM values are fairly similar across deciles within each subject, with the lowest values generally observed among students in the middle score deciles and somewhat larger values observed at the lower and upper ends of the score distribution. In Math, CSEM values are relatively stable across deciles, with only modestly higher variability in the lowest decile. In Reading, the same general pattern is evident, although the spread is somewhat wider in the lowest deciles, especially the first decile. Overall, the figure indicates that measurement precision is strongest for students in the middle of the score distribution and somewhat lower for students at the extremes.

Figure 4.3. Boxplots of CSEM by Subject and RIT Score Deciles

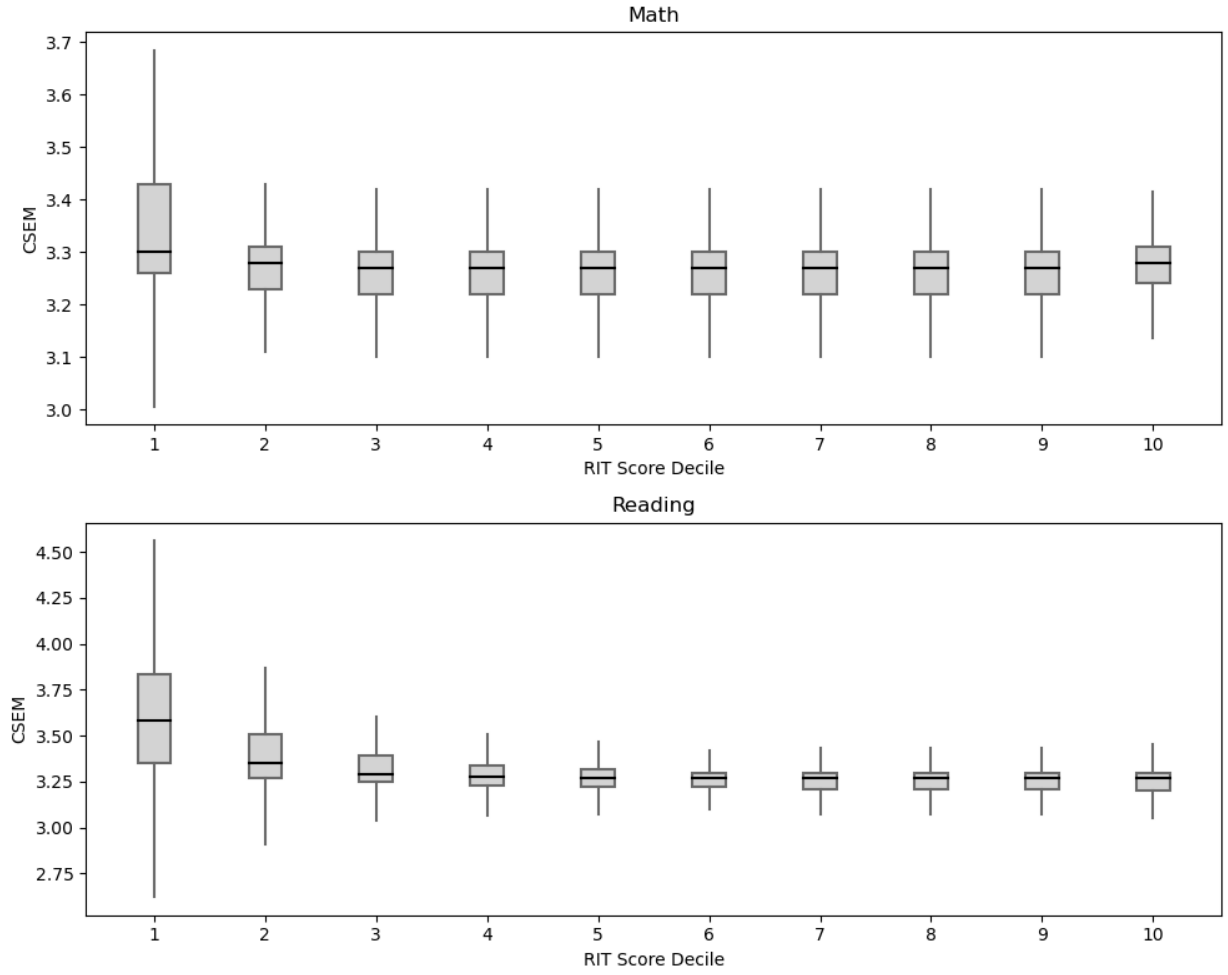


Table 4.5 presents numerical summaries of the CSEM values shown in Figure 4.3 by subject, grade, score group, and term for 2024–2025. Across subjects, grades, and terms, average CSEM values are generally quite similar. In most cases, the lowest CSEM values occur for students in the middle score group (5th decile), while somewhat higher values are observed for students in the lowest and highest score groups (1st and 10th deciles). This pattern is especially pronounced in Reading, where first-decile CSEM values are consistently higher across grades, though it is also evident in Math, particularly in the upper grades. Overall, the results show that differences in average CSEM values across grades and terms are modest, supporting that score precision is generally consistent across administrations.

Table 4.5. CSEM Summary by Subject, Grade, and Decile Score Group

Subject	Grade	Fall Score Groups			Winter Score Groups			Spring Score Groups		
		1st	5th	10th	1st	5th	10th	1st	5th	10th
Math	K	3.41	3.28	3.37	3.39	3.29	3.34	3.41	3.29	3.32
Math	1	3.32	3.28	3.30	3.31	3.30	3.31	3.31	3.30	3.31
Math	2	3.35	3.25	3.28	3.37	3.25	3.28	3.41	3.26	3.28
Math	3	3.35	3.24	3.28	3.32	3.25	3.25	3.34	3.26	3.25
Math	4	3.36	3.26	3.28	3.31	3.26	3.27	3.31	3.26	3.27

Subject	Grade	Fall Score Groups			Winter Score Groups			Spring Score Groups		
		1st	5th	10th	1st	5th	10th	1st	5th	10th
Math	5	3.36	3.26	3.26	3.32	3.26	3.28	3.32	3.26	3.29
Math	6	3.58	3.31	3.29	3.61	3.27	3.31	3.68	3.29	3.29
Math	7	3.64	3.31	3.29	3.61	3.28	3.33	3.60	3.29	3.36
Math	8	3.63	3.33	3.31	3.59	3.27	3.42	3.62	3.31	3.40
Math	HS	3.59	3.35	3.49	3.54	3.30	3.66	3.59	3.32	3.58
Reading	K	3.85	3.28	3.42	3.85	3.29	3.35	3.84	3.28	3.33
Reading	1	3.40	3.29	3.30	3.40	3.29	3.30	3.38	3.29	3.31
Reading	2	3.62	3.37	3.34	3.56	3.38	3.26	3.57	3.32	3.26
Reading	3	3.83	3.27	3.33	3.80	3.25	3.24	3.77	3.23	3.23
Reading	4	3.68	3.27	3.31	3.65	3.24	3.23	3.59	3.24	3.22
Reading	5	3.57	3.28	3.32	3.55	3.24	3.25	3.49	3.24	3.24
Reading	6	3.76	3.31	3.31	3.74	3.26	3.29	3.69	3.25	3.30
Reading	7	3.69	3.32	3.30	3.70	3.28	3.29	3.66	3.26	3.29
Reading	8	3.63	3.32	3.29	3.60	3.30	3.29	3.60	3.27	3.31
Reading	HS	3.69	3.35	3.35	3.70	3.31	3.32	3.72	3.29	3.40

Table 4.6 presents average CSEM values by subject, demographic subgroup, score group, and term. Across demographic subgroups and administrations, average CSEM values are generally very similar within each subject. For both Math and Reading, the middle score group consistently shows the lowest CSEM values, while the low and, to a lesser extent, high score groups tend to show slightly larger values. This pattern is especially noticeable in Reading, where the low score group regularly has the largest CSEM values across subgroups and terms. Overall, the table indicates that score precision is broadly consistent across subgroup populations, with only modest variation by term and score group.

Table 4.6. CSEM Summary by Subject, Demographic Group, Term, and Decile Score Group

Subject	Demographic Group	Fall Score Groups			Winter Score Groups			Spring Score Groups		
		Low	Middle	High	Low	Middle	High	Low	Middle	High
Math	Female	3.37	3.27	3.34	3.37	3.27	3.33	3.40	3.28	3.31
Math	Male	3.40	3.27	3.32	3.39	3.28	3.31	3.43	3.28	3.30
Math	White	3.38	3.28	3.33	3.38	3.28	3.32	3.43	3.28	3.31
Math	Black	3.36	3.27	3.31	3.37	3.28	3.33	3.39	3.29	3.33
Math	Asian	3.40	3.28	3.30	3.39	3.27	3.33	3.38	3.27	3.31
Math	Hispanic	3.39	3.27	3.32	3.38	3.27	3.32	3.42	3.28	3.31
Math	AI/AN	3.40	3.26	3.37	3.36	3.27	3.33	3.44	3.28	3.31
Math	NH/PI	3.35	3.26	--	3.36	3.29	3.30	3.40	3.28	3.31
Math	Multiple	3.36	3.25	3.33	3.36	3.27	3.29	3.35	3.26	3.33
Math	Other	3.39	3.30	3.31	3.37	3.29	3.31	3.35	3.26	3.30
Reading	Female	3.66	3.30	3.32	3.63	3.28	3.27	3.60	3.27	3.27
Reading	Male	3.67	3.30	3.32	3.64	3.29	3.27	3.61	3.27	3.27
Reading	White	3.64	3.29	3.31	3.58	3.29	3.27	3.55	3.27	3.28
Reading	Black	3.63	3.32	3.37	3.59	3.29	3.29	3.56	3.27	3.32
Reading	Asian	3.59	3.30	3.33	3.56	3.29	3.27	3.64	3.27	3.33

Subject	Demographic Group	Fall Score Groups			Winter Score Groups			Spring Score Groups		
		Low	Middle	High	Low	Middle	High	Low	Middle	High
Reading	Hispanic	3.67	3.30	3.32	3.64	3.29	3.27	3.61	3.27	3.27
Reading	AI/AN	3.72	3.30	3.33	3.70	3.29	3.25	3.64	3.26	3.26
Reading	NH/PI	3.68	3.31	3.33	3.62	3.28	3.24	3.62	3.26	3.31
Reading	Multiple	3.68	3.29	3.30	3.66	3.29	3.27	3.63	3.27	3.28
Reading	Other	3.62	3.30	3.29	3.60	3.29	3.28	3.63	3.27	3.28

4.6. Score Reliability

Reliability provides a scale-free summary of score precision for a group of examinees. Unlike the CSEM, which is reported for individual students and depends on the score scale, reliability summarizes the consistency of scores at the group level and is not affected by changes in scale. In item response theory, marginal reliability is commonly used as an index of internal consistency, combining information about score variance and measurement error across examinees. Higher reliability values indicate more consistent scores and less measurement error. In this addendum, 2024–2025 reliability estimates are reported across terms to evaluate the stability and consistency of scores over time.

4.6.1. Marginal Reliability

Marginal reliability estimates for MAP Growth Spanish Math and Spanish Reading were examined across terms (fall, winter, spring) and grades. Results indicate consistently strong score reliability across both subjects and all administrations, with reliability estimates generally meeting or exceeding commonly accepted thresholds for educational decision-making.

At the subject level, marginal reliabilities are uniformly high across terms, as shown in Table 4.7. In Math, reliability values range from 0.89 to 0.98, while in Reading, values range from 0.84 to 0.97. Reliability generally increases across grades within each subject, with the lowest estimates observed in kindergarten and the highest in the upper grades.

Across terms, marginal reliability is generally stable, with a slight tendency for winter and spring administrations to yield higher reliability estimates than fall, particularly in the early grades. This pattern is especially evident in kindergarten Reading, where reliability increases from 0.84 in fall to 0.91 in winter and to 0.94 in spring and in kindergarten Math, where values rise from 0.89 to 0.92 to 0.94. Although reliability increases modestly over time, differences across terms are small overall and do not meaningfully affect score interpretation.

Table 4.7. Marginal Reliability by Subject, Grade, and Term

Subject	Grade	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Math	K	28,648	0.89	29,666	0.92	30,903	0.94
Math	1	31,683	0.93	31,375	0.93	32,140	0.95
Math	2	25,837	0.94	24,558	0.95	25,436	0.95
Math	3	21,141	0.95	20,205	0.95	20,763	0.96
Math	4	14,545	0.96	13,602	0.96	14,109	0.97
Math	5	11,341	0.96	10,798	0.96	11,105	0.97
Math	6	4,809	0.95	4,246	0.95	4,736	0.96
Math	7	4,807	0.95	4,278	0.95	4,612	0.97

Subject	Grade	Fall		Winter		Spring	
		N	Reliability	N	Reliability	N	Reliability
Math	8	4,323	0.96	3,578	0.96	4,102	0.97
Math	HS	9,370	0.97	7,975	0.97	8,193	0.98
Reading	K	29,172	0.84	30,693	0.91	32,105	0.94
Reading	1	34,587	0.93	34,354	0.94	35,351	0.95
Reading	2	40,139	0.93	40,688	0.94	41,699	0.95
Reading	3	41,157	0.94	39,777	0.95	39,981	0.95
Reading	4	35,672	0.95	34,245	0.95	34,523	0.96
Reading	5	28,934	0.96	27,777	0.96	27,461	0.96
Reading	6	8,781	0.95	7,657	0.95	7,708	0.95
Reading	7	7,567	0.95	6,638	0.95	6,670	0.96
Reading	8	6,660	0.96	5,465	0.96	5,808	0.96
Reading	HS	2,315	0.96	1,950	0.96	1,702	0.97

4.6.2. Test-Retest Reliability

Test-retest reliability with alternate forms was examined for MAP Growth Spanish Math and Spanish Reading across consecutive testing windows within the 2024–2025 academic year: fall–winter, fall–spring, and winter–spring. Overall, the results presented in Table 4.8 indicate moderate to strong stability across grades and subjects, with a consistent pattern in which winter–spring reliabilities are highest, fall–winter reliabilities are slightly lower, and fall–spring reliabilities are lowest. This pattern is expected because the fall–spring interval spans more time for true academic growth to occur, which reduces stability coefficients.

In Math, reliability generally increases from kindergarten through grade 5, with the strongest coefficients observed in the elementary and middle grades. Winter–spring reliabilities range from 0.73 in kindergarten to 0.90 in grade 5, while fall–spring values are lower, ranging from 0.58 to 0.86 across grades. Reliability declines somewhat in the upper grades, particularly in grade 8 and high school, where fall–spring coefficients drop to 0.69 and 0.52, respectively. Kindergarten Math shows the lowest reliability estimates overall, consistent with greater developmental variability at the earliest grade levels.

Reading shows a similar pattern, with reliability improving substantially from kindergarten into the elementary grades and remaining generally strong through middle school. Winter–spring coefficients range from 0.71 in kindergarten to 0.88 in grade 5, while fall–spring coefficients are again the lowest across nearly all grades, ranging from 0.56 to 0.84. Reading reliability remains relatively stable in grades 3 through 8, though somewhat lower values appear in high school, especially for the fall–spring comparison (0.78). As in Math, kindergarten Reading shows the lowest stability coefficients, reflecting more rapid skill development and greater measurement error in early literacy.

Table 4.8. Test-Retest Reliability by Subject, Grade, and Term

Subject	Grade	N	Fall/Winter	Fall/Spring	Winter/Spring
			Reliability	Reliability	Reliability
Math	K	22,071	0.68	0.58	0.73
Math	1	24,870	0.79	0.74	0.81
Math	2	18,065	0.83	0.79	0.85

Subject	Grade	N	Fall/Winter Reliability	Fall/Spring Reliability	Winter/Spring Reliability
Math	3	14,615	0.86	0.82	0.87
Math	4	9,354	0.88	0.85	0.89
Math	5	6,715	0.90	0.86	0.90
Math	6	1,968	0.83	0.80	0.85
Math	7	1,968	0.81	0.71	0.76
Math	8	1,571	0.77	0.69	0.75
Math	HS	2,372	0.68	0.52	0.64
Reading	K	24,191	0.65	0.56	0.71
Reading	1	28,823	0.80	0.74	0.82
Reading	2	33,806	0.82	0.77	0.84
Reading	3	32,865	0.83	0.80	0.86
Reading	4	27,891	0.85	0.83	0.87
Reading	5	21,673	0.86	0.84	0.88
Reading	6	4,556	0.84	0.82	0.84
Reading	7	3,466	0.85	0.81	0.84
Reading	8	2,840	0.86	0.82	0.85
Reading	HS	293	0.88	0.78	0.82

5. The Test Is Fair for All Examinees

5.1. Test Taker Demographics

Students taking the MAP Growth Spanish Math and Reading tests in 2024–2025 were about equally distributed between male and female, as shown in Table 5.1. Examinees were largely Hispanic and represent about 80% of examinees in every term, as shown in Table 5.2. White students represent the second-largest demographic with 7–10% of examinees. The third- and fourth-largest groups represent either multiple races or did not specify their race and ethnicity.

Table 5.1. Demographics by Gender

Subject	Term	Total	Percentage Male	Percentage Female
Math	Fall	156,406	50.09	49.91
Math	Winter	150,167	50.07	49.93
Math	Spring	155,982	49.89	50.11
Reading	Fall	234,963	49.79	50.21
Reading	Winter	229,221	49.80	50.20
Reading	Spring	232,993	49.70	50.30

Table 5.2. Demographics by Race/Ethnicity

Subject	Term	Total	Percentage							
			Asian	Black	NH/PI	AI/AN	Hispanic	White	Other	Multiple
Math	Fall	156,406	0.28	1.13	0.13	1.71	84.20	8.05	1.85	2.65
Math	Winter	150,167	0.24	1.09	0.15	1.61	84.78	7.84	2.26	2.04
Math	Spring	155,982	0.25	1.07	0.12	1.36	82.00	7.51	4.58	3.11
Reading	Fall	234,963	0.46	1.68	0.15	1.87	79.48	11.68	1.61	3.07
Reading	Winter	229,221	0.41	1.68	0.14	1.83	80.71	10.80	1.86	2.57
Reading	Spring	232,993	0.44	1.57	0.12	1.33	78.20	10.26	3.20	4.89

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

5.2. Differential Item Functioning

Students taking the Spanish MAP Growth tests in 2024–2025 represented a fairly homogenous population, as indicated in the student demographics shown in Table 5.2. This demographic composition of the Spanish test takers limits the number of group comparison that are possible when evaluating differential item functioning (DIF). As shown in Table 5.3, the DIF analysis compared male and female examinees and compared Hispanic and White examinees. Results indicate the vast majority (91–97%) of items are classified as “A” items (i.e., combined A+ and A- items), which indicates no DIF or negligible amounts of DIF. Another 3–6% of items showed moderate levels of DIF and were classified as “B” items. Few items were classified as showing large DIF, or “C” items, with percentages ranging from 0.5 to 2. Overall results indicate that items functioning similarly for different groups of examinees taking the Spanish MAP Growth tests.

Table 5.3. DIF Classification Summary

Subject	Term	Focal Group	Reference Group	N Items	Percentage in DIF Class					
					A+	A-	B+	B-	C+	C-
Math	Fall	Female	Male	17,720	50.71	40.63	1.61	4.99	0.13	1.92
Math	Fall	Hispanic	White	16,893	44.35	48.46	2.50	3.36	0.26	1.07
Math	Winter	Female	Male	18,016	51.72	40.36	1.48	4.60	0.13	1.71
Math	Winter	Hispanic	White	17,305	46.31	46.93	2.12	3.06	0.46	1.11
Math	Spring	Female	Male	18,552	52.93	39.81	1.54	4.16	0.09	1.48
Math	Spring	Hispanic	White	17,836	48.12	45.43	2.07	2.75	0.66	0.98
Reading	Fall	Female	Male	11,070	56.03	40.88	1.13	1.41	0.11	0.45
Reading	Fall	Hispanic	White	10,921	42.62	53.85	0.42	2.48	0.03	0.60
Reading	Winter	Female	Male	11,135	57.82	39.18	1.17	1.31	0.06	0.46
Reading	Winter	Hispanic	White	10,956	41.21	54.72	0.43	2.85	0.03	0.77
Reading	Spring	Female	Male	11,151	58.36	38.76	1.00	1.38	0.08	0.41
Reading	Spring	Hispanic	White	10,964	41.87	54.30	0.37	2.71	0.05	0.68

References

Achieve, Inc. (2019). *A framework to evaluate cognitive complexity in mathematics assessments*. https://www.achieve.org/files/Mathematics%20Cognitive%20Complexity%20Framework_Final_92619.pdf

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

Jiban, C. (2017). *MAP Growth reading and language usage literature review*. NWEA.

NWEA (2026). *MAP Growth technical report for 2024–2025*. NWEA.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/onlinepubs/synth44.pdf

Wang, S., & Li, S. (2020). *Calibration of Spanish MAP Growth math tests*. NWEA. https://www.nwea.org/uploads/2021/06/Spanish-MAP-Growth-Math-Calibration-2020-06-15_NWEA_report.pdf