RENAISSANCE®

Star Assessments™ for Spanish–Reading Technical Documentation

RENAISSANCE Star Spanish®

RENAISSANCE Star Reading®

Renaissance Learning
PO Box 8036
Wisconsin Rapids, WI 54495-8036
Telephone: (800) 338-4204
(715) 424-3636

Outside the US: 1.715.424.3636
Fax: (715) 424-4242
Email (general questions): answers@renaissance.com
Email (technical questions): support@renaissance.com
Email (international support): worldsupport@renaissance.com
Website: www.renaissance.com

Copyright Notice

6/2020 SRSp

Contents

Introduction

Star Reading Spanish: Screening and Progress-Monitoring Assessment

In the spring of 2011, Renaissance conducted a pilot study of Star Reading Spanish. The goal was to determine the feasibility of developing a vertical scale for Spanish reading content and for the intended Spanish-speaking population. Following a successful pilot phase, a much larger scale calibration phase began in December of 2011. The result of the calibration study was the publication of the first version of Star Reading Spanish in the fall of 2012. This version of Star Reading Spanish comprised only 25 vocabulary-in-context items administered adaptively in grades 1 to 6 only.

Due to a growing need to assess multiple standards and skills in Spanish reading comprehension, Renaissance embarked on the development of a second version of Star Reading Spanish in the fall of 2014.

Star Reading Spanish is a 34-item standards-based adaptive assessment, aligned to state and national curriculum standards, that takes an average of less than 21 minutes to complete. Star Reading Spanish provides immediate feedback to teachers and administrators on each student's Spanish reading development.

The test development of this version of Star Reading Spanish began with the translation of the Star Reading English test items into Spanish. Transadaptation is a process that begins with translation, but is followed by revision or replacement of text that, while it may be a valid test of student skills in English, would not be so in Spanish. Following transadaptation, a rigorous and lengthy process of pilot testing, item calibration and scale development, and still further research to fine-tune the initial item calibrations was conducted over several years. This resulted in the updated and current version of Star Reading Spanish which was released at the start of the 2018–2019 school year. Star Reading Spanish is designed to measure Spanish Reading achievement of students in grades K–12; however, the current version contains items for only grades K–8 with ongoing work to extend the full content coverage up to grade 12 in the future. All students in grade K–12 can test with Star Reading Spanish.

Star Reading Spanish Purpose

As a periodic progress-monitoring assessment, Star Reading Spanish progress monitoring serves three purposes. First, it provides educators with quick and accurate estimates of Spanish reading comprehension using students' instructional reading levels. Second, it assesses Spanish reading achievement relative to norms based on nationwide user data. Third, it provides the means for tracking growth in a consistent manner longitudinally for all students. This is especially helpful to school- and district-level administrators.

While the Star Reading Spanish test provides accurate normed data like traditional norm-referenced tests, it is not intended to be used as a "high-stakes" test. Generally, states are required to use high-stakes assessments to document growth, adequate yearly progress, and mastery of state standards. These high-stakes tests are also used to report end-of-period performance to parents and administrators or to determine eligibility for promotion or placement. Star Reading Spanish is not intended for these purposes. Rather, when a high correlation between the Star Reading Spanish test and high-stakes instruments exists, classroom teachers can use Star Reading Spanish scores to fine-tune instruction while there is still time to improve performance before the regular test cycle. Furthermore, Star Reading Spanish results can easily be disaggregated to identify and address the needs of various groups of students.

The Star Reading Spanish test's repeatability and flexible administration provide specific advantages for everyone responsible for the education process:

- For students, Star Reading Spanish software provides a challenging, interactive, and brief test that builds confidence in their Spanish reading ability.
- For teachers, the Star Reading Spanish test facilitates individualized instruction by identifying children who need remediation or enrichment most.
- For principals, the Star Reading Spanish software provides regular, accurate reports on performance at the class, grade, building, and district level.
- For district administrators and assessment specialists, it provides a wealth of reliable and timely data on reading growth at each school and districtwide. It also provides a valid basis for comparing data across schools, grades, and special student populations.

This technical report documents the suitability of the Star Reading Spanish computer-adaptive test for these purposes and demonstrates quantitatively how well this innovative instrument in Spanish reading assessment performs.

Design of Star Reading Spanish

Two Generations of Star Reading Spanish Assessments

The introduction of Star Reading Spanish in 2012 marked the first generation of Star Reading Spanish assessments. That version was adaptive and built on the item response theory (IRT) framework. The item bank consisted of 850+ vocabulary-in-context items. Despite being a breakthrough assessment in Spanish reading, it only assessed reading using a single item type.

The second generation is represented by the current version of Star Reading Spanish, published in 2018.

This version of Star Reading Spanish is designed as a standards-based test; its items are organized into 5 blueprint domains, 11 skill sets, 35 general skills, and over 300 discrete skills—all designed to align to national and state curriculum standards in reading and language arts, including state-specific as well as the Common Core State Standards. Star Reading Spanish uses fixed-length adaptive tests—34 items in length—both to facilitate broad standards coverage and to maintain high measurement precision and reliability.

Overarching Design Considerations

One of the fundamental Star Reading Spanish design decisions involved the choice of how to administer the test. The primary advantage of using computer software to administer Star Reading Spanish tests is the ability to tailor each student's test based on his or her responses to previous items.

Conventional assessments, including paper-and-pencil tests, typically entail fixed test forms: every student must respond to the same items in the same sequence. Using computer-adaptive procedures makes it possible for students to test on items that appropriately match their current level of proficiency. The item selection procedures, termed Adaptive Branching, effectively customize the test for each student's achievement level.

Adaptive Branching offers significant advantages in terms of test reliability, testing time, and student motivation. Reliability improves over fixed-form tests because the test difficulty is adjusted to each individual's performance level; students do not have to fit a "one test fits all" model. Most of the test items that students respond to are at levels of difficulty that closely match their

achievement level. Testing time decreases because, unlike in paper-and-pencil tests, there is no need to expose every student to a broad range of material, portions of which are inappropriate because they are either too easy for high achievers or too difficult for those with low current levels of performance. Finally, student motivation improves simply because of these issues—test time is minimized and test content is neither too difficult nor too easy.

Another fundamental Star Reading Spanish design decision involved the choice of the content and format of items for the test. Many types of stimulus and response procedures were explored, researched, discussed, and prototyped using the Star Reading English items. These procedures included the traditional reading passage followed by sets of literal or inferential questions, previously published extended selections of text followed by open-ended questions requiring student-constructed answers, and several cloze-type procedures for passage presentation. While all of these procedures can be used to measure reading comprehension and overall reading achievement, the vocabulary-in-context format was selected as the primary item format for the first-generation Star Reading Spanish assessments. This decision was made for interrelated reasons of efficiency, breadth of construct coverage, and objectivity and simplicity of scoring.

Each Star Reading Spanish test consists of 34 items; of these, the first 10 are vocabulary-in-context items, while the last 24 items spiral their content to include standards-based material from all five blueprint domains.

For these reasons, the Star Reading Spanish test design and item formats provide a valid procedure for assessing a student's reading comprehension in Spanish. Data and information presented in this document reinforce this.

Star Reading Spanish is designed specifically for use on a computer with web access. All management and test administration functions are controlled using a management system which is accessed by means of a computer with web access.

This makes a number of new features possible:

- It makes it possible for multiple schools to share a central database, such as a district-level database. Records of students transferring between schools within the district will be maintained in the database; the only information that needs revision following a transfer is the student's updated school and class assignments.
- The same database that contains Star Reading Spanish data can contain data on other Star tests, including Star Early Literacy and Star Math, in both English and Spanish. The Renaissance program is a powerful information management program that allows you to manage all your

district, school, personnel, parent, and student data in one place. Changes made to district, school, teacher, parent, and student data for any of these products, as well as other Renaissance software, are reflected in every other Renaissance program sharing the central database.

- Multiple levels of access are available, from the test administrator within a school or classroom to teachers, principals, district administrators, and even parents.
- Renaissance takes reporting to a new level. Not only can you generate reports from the student level all the way up to the school level, but you can also limit reports to specific groups, subgroups, and combinations of subgroups. This supports "disaggregated" reporting; for example, a report might be specific to students eligible for free or reduced lunch, to English language learners, or to students who fit both categories. It also supports compiling reports by teacher, class, school, grade within a school, and many other criteria such as a specific date range. In addition, the Renaissance consolidated reports allow you to gather data from more than one program (such as Star Reading and Accelerated Reader) at the teacher, class, school, and district level and display the information in one report.
- Since the Renaissance software is accessed through a web browser, teachers (and administrators) will be able to access the program from home—provided the district or school gives them that access.

Test Interface

The Star Reading test interface was designed to be both simple and effective. All test questions consist of either a single sentence or a full paragraph, followed by three or four response alternatives. Students can use either the mouse or the keyboard to answer questions.

- If using the keyboard, students press one of the four number keys (**1**, **2**, **3**, or **4**) and then press the **Enter** key (or the **return** key on Macintosh computers).
- If using the mouse, students click the answer of choice and then click **Siguiente** to enter the answer.
- On a tablet, students tap their answer choice; then, they tap **Siguiente**.

Practice Session

Star Reading Spanish software includes a provision for a brief practice test preceding the test itself. The practice session allows students to get comfortable with the test interface and to make sure that they know how to operate it properly. As soon as a student has answered three practice questions correctly, the program takes the student into the actual test. Even the lowest-level readers should be able to answer the sample questions correctly. If the student has not successfully answered three items by the end of the practice session, Star Reading Spanish will halt the testing session and tell the student to ask the teacher for help. It may be that the student cannot read at even the most basic level, or it may be that the student needs help operating the interface, in which case the teacher should help the student through the practice session the next time. Before beginning the next test with the student, the program will recommend that the teacher assist the student during the practice.

Once a student has successfully passed a practice session, the student will not be presented with practice items again on a test of the same type taken within the next 180 days.

Adaptive Branching/Test Length

Star Reading's Spanish branching control uses a proprietary approach somewhat more complex than the simple Rasch maximum information IRT model. The Star Reading Spanish approach was designed to yield reliable test results for both the criterion-referenced and norm-referenced scores by adjusting item difficulty to the responses of the individual being tested while striving to minimize test length and student frustration.

In order to minimize student frustration, the first administration of the Star Reading Spanish test begins with items that have a difficulty level that is below what a typical student at a given grade can handle—usually one or two grades below grade placement. On the average, about 85 percent of students will be able to answer the first item correctly. Teachers can override this typical value by entering an even lower Estimated Instructional Reading Level for the student.

After the first two items, Star Reading Spanish strikes a balance between student motivation and measurement efficiency by tailoring the choice of test items such that students answer an average of 67 percent of items correctly. On the second and subsequent administrations, the Star Reading test

again begins with items that have a difficulty level lower than the previously demonstrated reading ability. Students generally have an 85 percent chance of answering the first item correctly on second and subsequent tests.

Test Length

Once the testing session is underway, the Star Reading Spanish test administers 34 items of varying difficulty based on the student's responses; this is sufficient information to obtain a reliable Scaled Score and to determine the student's Instructional Reading Level.

The length of time needed to complete a Star Reading Spanish test varies across students. Table 1 provides an overview of the testing time by grade for the students who took Star Reading Spanish during the 2016–2017 school year. The results of the analysis of test completion time indicate that half or more of students completed the test in less than 21 minutes, depending on grade, and at least 95% of students at every grade finished their Star Reading Spanish test in less than 36 minutes.

Table 1: Average and Percentiles of Total Time to Complete the Star Reading Spanish Assessment During the 2016–2017 School Year

Grade	Sample Size	Time to Complete Test (in Minutes)					
		Mean	Standard Deviation	5th Percentile	50th Percentile	95th Percentile	99th Percentile
1	25,436	17.49	9.50	5.57	15.67	35.63	44.54
2	54,970	18.28	8.43	6.38	17.30	33.63	41.53
3	54,175	18.43	6.80	7.58	18.23	29.88	35.54
4	46,267	20.00	6.73	8.58	20.12	31.00	35.83
5	36,209	19.53	6.55	8.27	19.67	30.15	34.57
6	13,007	20.11	6.36	9.08	20.15	30.35	35.23
7	8,534	19.52	6.30	8.43	19.78	29.72	33.62
8	8,628	19.98	6.10	9.30	20.27	29.66	34.00

Test Repetition

Star Reading Spanish score data can be used for multiple purposes such as screening, placement, planning instruction, benchmarking, and outcomes measurement. The frequency with which the assessment is administered depends on the purpose for assessment and how the data will be used.

Renaissance recommends assessing students only as frequently as necessary to get the data needed. Schools that use Star for screening purposes typically administer it two to five times per year. Star Reading Spanish may be administered three times a year for progress monitoring purposes.

Star Reading Spanish keeps track of the questions presented to each student from test session to test session and will not ask the same question more than once in any 90-day period.

Item Time Limits

Star Reading Spanish tests place no limits on total testing time. However, there are time limits for each test item. The per-item time limits are generous and ensure that more than 90 percent of students can complete each item within the normal time limits.

Star Reading Spanish provides the option of extended time limits for selected students who, in the judgment of the test administrator, require more than the standard amount of time to read and answer the test questions.

Extended time may be a valuable accommodation for Spanish language learners as well as for some students with disabilities. Test users who elect the extended time limit for their students should be aware that Star Reading Spanish norms, as well as other technical data such as reliability and validity, are based on test administration using the standard time limits. When the extended time limit accommodation is elected, students have longer than the standard time limits to answer each question.

Table 2 shows the Star Reading Spanish test time-out limits for individual items. These time limits are based on a student's grade level.

Table 2: Star Reading Spanish Time-Out Limits

Grade	Question Type	Standard Time Limit (seconds/item)	Extended Time Limit (seconds/item)
K–2	Practice	120	300
	Test, all questions (operational and uncalibrated items)	120	300
3–12	Practice	120	300
	Test, all questions (operational and uncalibrated items)	90	180

These time-out values are based on latency data obtained during item calibration.

At all grades, regardless of the extended time limit setting, when a student has only 15 seconds remaining for a given item, a time-out warning appears, indicating that he or she should make a final selection and move on. Items that time out are counted as incorrect responses *unless* the student has the correct answer selected when the item times out. If the correct answer has been selected at that time, the item will be counted as a correct response.

If a student doesn't respond to an item, the item times out and briefly gives the student a message describing what has happened. Then the next item is presented. The student does not have an opportunity to take the item again. If a student doesn't respond to any item, all items are scored as incorrect.

Test Security

Star Reading Spanish software includes a number of security features to protect the content of the test and to maintain the confidentiality of the test results.

Split-Application Model

When students log into Star Reading Spanish, they do not have access to the same functions that teachers, administrators, and other personnel can access. Students are allowed to take the test, but no other features available in Star Reading Spanish are available to them; therefore, they have no access to confidential information. When teachers and administrators log in, they can manage student and class information, set preferences, and create informative reports about student test performance.

Individualized Tests

Using Adaptive Branching, every Star Reading Spanish test consists of items chosen from a large number of items of similar difficulty based on the student's estimated ability. Because each test is individually assembled based on the student's past and present performance, identical sequences of items are rare. This feature, while motivated chiefly by psychometric considerations, contributes to test security by limiting the impact of item exposure.

Data Encryption

A major defense against unauthorized access to test content and student test scores is data encryption. All of the items and export files are encrypted. Without the appropriate decryption code, it is practically impossible to read the Star Reading Spanish data or access or change it with other software.

Access Levels and Capabilities

Each user's level of access to a Renaissance program depends on the primary position assigned to that user. Each primary position is part of a group: these groups have different names depending on the which platform the user's Renaissance site is on.

- For customers on the original platform, the groups are called *user groups,* and there are seven of them (district administrator, district staff, school administrator, school staff, teachers, students, and parents). Each user group is granted a specific set of *capabilities*.
- For customers who have been migrated to the new Renaissance Growth Platform, the groups are called *user permission groups,* and there are six of them (district level administrator, district dashboard owner, district staff, school level administrator, school staff, and teacher). Each user permission group is granted a specific set of *user permissions.*

Each capability or user permission corresponds to one or more tasks that can be performed in the program. The capabilities/user permissions for these groups can be changed, and they can be granted or removed on an individual level.

Renaissance also allows you to restrict students' access to certain computers. This prevents students from taking Star Reading Spanish tests from unauthorized computers (such as home computers). For more information, see https://help.renaissance.com/RP/SettingSecurityOptions or https://help2.renaissance.com/setup/22509.

The security of the Star Reading Spanish data is also protected by each person's user name (which must be unique) and password. User names and passwords identify users, and the program only allows them access to the data and features that they are allowed based on their primary position and the user permissions that they have been granted. Personnel who log in to Renaissance (teachers, administrators, or staff) must enter a user name and password before they can access the data and create reports. Parents on original sites who are granted access to Renaissance must also log in with a user name and password before they can access information about their

children. Without an appropriate user name and password, personnel and parents cannot use the Star Reading Spanish software.

Test Monitoring/Password Entry

Test monitoring is another useful Star Reading Spanish security feature. Test monitoring is implemented using the Password Requirement preference, which specifies whether monitors must enter their passwords at the start of a test. Students are required to enter a user name and password to log in before taking a test. This ensures that students cannot take tests using other students' names.

Final Caveat

While Star Reading Spanish software can do much to provide specific measures of test security, the most important line of defense against unauthorized access or misuse of the program is the user's responsibility. Teachers and test monitors need to be careful not to leave the program running unattended and to monitor all testing to prevent students from cheating, copying down questions and answers, or performing "print screens" during a test session. Taking these simple precautionary steps will help maintain Star Reading Spanish's security and the quality and validity of its scores.

Test Administration Procedures

In order to ensure consistency and comparability of results to the Star Reading Spanish norms, students taking Star Reading Spanish tests should follow standard administration procedures. The testing environment should be as free from distractions for the student as possible.

The Pretest Instructions included with the Star Reading Spanish product describes the standard test orientation procedures that teachers should follow to prepare their students for the Star Reading Spanish test. These instructions are intended for use with students of all ages. The instructions were successfully field-tested with students ranging from grades 1–8. It is important to use these same instructions with all students before they take the Star Reading Spanish test.

Content and Item Development

Content Specification: Star Reading Spanish

The scale and scope of Star Reading Spanish continues to grow since it was first released in 2012. It was initially an assessment consisting exclusively of Vocabulary-in-Context items; for the current version, other item types have been added to test additional reading skills. New items are commonly developed to maintain the integrity of the item bank and are added to the operational bank only after they are determined to be psychometrically valid.

Star Reading Spanish is based upon the assessment of 36 general skills organized within 5 Blueprint Domains of reading (see Table 3), and maps the progressions of reading skills and understandings as they develop in sophistication from kindergarten through grade 12. Each Star item is designed to assess a specific skill within the test blueprint. The Star Reading Spanish test blueprint is largely fixed. Renaissance may alter the blueprint if there are data-driven reasons to make a major change to the content.

For information regarding the development of Star Reading Spanish items, see “Item Development Specifications” on page 17. Before inclusion in the Star Reading Spanish item bank, all items are written to ensure they meet the content specifications for Star Reading Spanish item development and pass psychometric calibration. Items that do not meet the content and psychometric specifications are either discarded or revised for recalibration. All new item development adheres to the content specifications, and all items have been calibrated using the dynamic calibration method.

The first stage of expanded Star Reading Spanish development was to identify the set of skills to be assessed. The test design for Star Reading Spanish is identical to the English-language version of Star Reading. Differences between the two products appear primarily at the level of item-design for skills where there are linguistic and semantic differences between English and Spanish. Multiple resources were consulted to determine the set of skills most appropriate for assessing the reading development of K–12 US students. The resources include but are not limited to:

- *Reading Next—A Vision for Action and Research in Middle and High School Literacy: A Report to Carnegie Corporation of New York* © 2004 by Carnegie Corporation of New York. https://www.all4ed.org/wp-content/uploads/2006/07/ReadingNext.pdf.

- *NCTE Principles of Adolescent Literacy Reform, A Policy Research Brief*, Produced by The National Council of Teachers of English, April 2006. http://www.ncte.org/library/NCTEFiles/Resources/Positions/Adol-Lit-Brief.pdf.
- *Improving Adolescent Literacy: Effective Classroom and Intervention Practices*, August 2008. http://eric.ed.gov/PDFS/ED502398.pdf.
- *Reading Framework for the 2009 National Assessment of Education Progress*. http://www.nagb.org/publications/frameworks/reading09.pdf.
- Common Core State Standards Initiative (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*.
- *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects, Spanish Language Version*, San Diego County Office of Education. https://commoncore-espanol.sdcoe.net/CCSS-en-Espanol/SLA-Literacy
- Individual state standards from all 50 states.
- *Texas Essential Knowledge and Skills: Spanish Language Arts and Reading English as a Second Language*, Texas Education Agency. http://ritter.tea.state.tx.us/rules/tac/chapter128/index.html

The development of the skills list included iterative reviews by reading and assessment experts and psychometricians specializing in educational assessment. See Table 3 for the Star Reading Spanish Blueprint Skills List. The skills list is organized into five blueprint domains:

- Word Knowledge and Skills
- Comprehension Strategies and Constructing Meaning
- Analyzing Literary Text
- Understanding Author's Craft
- Analyzing Argument and Evaluating Text

The second stage of development was to develop and calibrate Spanish-language versions of psychometrically validated Star Reading English-language items. Star Reading Spanish items were drawn from English-language items using the process of transadaptation. Transadaptation involves both the translation and adaptation of English-language items and the replacement of items unfit for translation/transadaptation with items written in Spanish. This process ensures that test items accurately assess the targeted skills while also being sensitive to semantic differences between Spanish and English. All transadaptation was performed by a professional

Spanish-language translation vendor and reviewed by Spanish-fluent editors at Renaissance. A strict development process was maintained to ensure quality item development.

The third and ongoing stage of development is to develop items written directly in Spanish. All writers and editors have content-area expertise, relevant classroom experience, and native-level knowledge of Spanish and Spanish pedagogy, and they use those qualifications in determining grade-level appropriateness for each item developed. Grade-level appropriateness is determined by multiple factors including reading skill, reading level, cognitive load, vocabulary grade level, sentence structure, sentence length, subject matter, and interest level. A strict development process is maintained to ensure quality item development.

Assessment items, once written, edited, and reviewed, are field tested and psychometrically calibrated to estimate their Rasch difficulty parameters and goodness of fit to the model. Field testing and calibration are conducted in a single step. This dynamic calibration method is done by embedding new items in appropriate, random positions within the Star assessments to collect the item response data needed for psychometric evaluation and calibration analysis. Following these analyses, each assessment item—along with both traditional and Item Response Theory (IRT) analysis information (including fit plots) and information about the test level, form, and item identifier—is stored in an item statistics database. A panel of content reviewers then examines each item within the proper context to determine whether the item meets all criteria for use in an operational assessment.

Table 3: Star Reading Spanish Organization: Blueprint Domains, Skill Sets, and Skills

Star Reading Spanish Blueprint Domain	Star Reading Spanish Blueprint Skill Set	Star Reading Spanish Blueprint Skill
Word Knowledge and Skills	Vocabulary Strategies	Use context clues
		Use structural analysis
	Vocabulary Knowledge	Recognize and understand synonyms
		Recognize and understand homonyms and multi-meaning words
		Recognize connotation and denotation
		Understand idioms
		Understand analogies

Table 3: Star Reading Spanish Organization: Blueprint Domains, Skill Sets, and Skills

Star Reading Spanish Blueprint Domain	Star Reading Spanish Blueprint Skill Set	Star Reading Spanish Blueprint Skill
Comprehension Strategies and Constructing Meaning	Reading Process Skills	Make predictions
		Identify author's purpose
		Identify and understand text features
		Recognize an accurate summary of text
	Constructing Meaning	Understand vocabulary in context
		Draw conclusions
		Identify and understand main ideas
		Identify details
		Extend meaning and form generalizations
		Identify and differentiate fact and opinion
	Organizational Structure	Identify organizational structure
		Understand cause and effect
		Understand comparison and contrast
		Identify and understand sequence
Analyzing Literary Text	Literary Elements	Identify and understand elements of plot
		Identify and understand setting
		Identify characters and understand characterization
		Identify and understand theme
		Identify the narrator and point of view
	Genre Characteristics	Identify fiction and nonfiction, reality and fantasy
		Identify and understand characteristics of genres
Understanding Author's Craft	Author's Choices	Understand figurative language
		Understand literary devices
		Identify sensory detail
Analyzing Argument and Evaluating Text	Analysis	Identify bias and analyze text for logical fallacies
		Identify and understand persuasion
	Evaluation	Evaluate reasoning and support
		Evaluate credibility

Table 4: Example of Star Reading Spanish Item Adherence to a Specific Skill within Star Reading Blueprint Structure

<table>
<tr><td colspan="3">Blueprint Domain: Analyzing Literary Text</td></tr>
<tr><td colspan="3">Blueprint Skill Set: Literary Elements</td></tr>
<tr><td colspan="3">Blueprint Skill: Identify characters and understand characterization</td></tr>
<tr><td rowspan="3">Grade-level subskill statements:</td><td>2nd grade</td><td>Describe major and minor characters and their traits using key details.</td></tr>
<tr><td>3rd grade</td><td>Identify and describe main character's traits, motives, and feelings.
3rd Grade Star Reading Spanish Item
Daniela le había dicho a todos sus amigos que estaba entusiasmada por viajar a otro país. Su tía vivía en Canadá, a solo cuatro horas en avión. Para Daniela, sin embargo, parecía que estaba al otro lado del mundo. Iba a ser su primer viaje lejos de casa sin sus padres. Aunque sentía que sería una aventura, Daniela estaba un poco nerviosa.
¿ Cómo se siente Daniela por su viaje?
1. demasiado asustada para ir
2. entusiasmada y nerviosa
3. cansada de viajar</td></tr>
<tr><td>4th grade</td><td>Describe characters, interactions with other characters, and relationship between actions, traits, and motives.</td></tr>
</table>

Development of the Spanish Graded Vocabulary List

The original point of reference for the development of Star Reading Spanish items was the creation of a Spanish graded vocabulary list. Because no published graded vocabulary lists existed in Spanish, Renaissance linguistic experts decided to build a Spanish list by starting with the English graded vocabulary list from the ATOS readability formula. ATOS is a system for evaluating the reading level of continuous text; at the time of development of the Spanish Graded Vocabulary List it contained over 9,000 words in its graded-vocabulary list. This readability formula was developed by Renaissance and designed by leading readability experts. ATOS is the first formula to include statistics from actual student book reading.

Once the ATOS vocabulary list was selected as the basis for the Spanish Graded Vocabulary List, the words were translated and assigned grade levels by external Spanish-language linguistic experts and reviewed by Renaissance

prior to a pilot validation study conducted in 2011 and a formal review by The Center for Applied Linguistics.

The Spanish Graded Vocabulary List served as the basis for Star Reading Spanish's grades K–8 Vocabulary-in-Context items. The list is added to and refined on an ongoing basis.

Item Development Specifications

The Content item bank for Star Reading Spanish has been expanding steadily since the product launch and continues to this day. Content development is driven by the test design and test purposes, which are to measure comprehension and general reading achievement in Spanish. Based on test purpose, the desired content had to meet certain criteria. First, it had to cover a range broad enough to test students from grades K–12. Thus, items had to represent reading levels ranging all the way from kindergarten through post-high school. Second, the current collection of test items must be large enough so that students could test often without being given the same item twice. As of November 2018, the item bank for Star Reading Spanish contains over 2,200 items.

During item development, every effort is made to avoid the use of stereotypes, potentially offensive language or characterizations, and descriptions of people or events that could be construed as being offensive, demeaning, patronizing, or otherwise insensitive. The editing process also includes strict reviews for factual accuracy and for bias and sensitivity to attend to issues of gender and ethnic-group balance and fairness.

Vocabulary-in-Context Item Specifications

Vocabulary-in-context items are single-sentence measures of comprehension presented in a modified "cloze" format. Each vocabulary-in-context item is written to the following specifications:

1. Each vocabulary-in-context test item consists of a single-context sentence. This sentence contains a blank indicating a missing word. Three or four possible answers are shown beneath the sentence. For questions developed at a kindergarten or first-grade reading level, three possible answers are given. Questions at a second-grade reading level and higher offer four possible answers.
2. To answer the question, the student selects the word from the answer choices that best completes the sentence. The correct answer option is

the word that appropriately fits both the semantics and the syntax of the sentence. All of the incorrect answer options either fit the syntax of the sentence or relate to the meaning of something in the sentence. They do not, however, meet both conditions.

3. The answer blanks are generally located near the end of the context sentence to minimize the amount of rereading required.
4. The sentence provides sufficient context clues for students to determine the appropriate answer choice. However, the length of each sentence varies according to the guidelines shown in Table 5.

Table 5: Maximum Sentence Length per Item Grade Level

Item Grade Level	Maximum Sentence Length (Including Sentence Blank)
Kindergarten–Grade 1	12 words
Grades 2–3	14 words
Grades 4–6	16 words
Grades 7–8	18 words

5. Typically, the words that provide the context clues in the sentence are below the level of the actual test word. However, due to a limited number of available words, not all of the questions at or below grade 2 meet this criterion—but even at these levels, no context words are above the grade level of the item.
6. The correct answer option is a word selected from the appropriate grade level of the item set. Through vocabulary-in-context test items, Star Reading Spanish requires students to rely on background information, apply vocabulary knowledge, and use active strategies to construct meaning from the assessment text. These cognitive tasks are consistent with what researchers and practitioners describe as reading comprehension.

Reading Skill Item Specifications

Star Reading Spanish maps the progression of reading skills and understandings as they develop in sophistication from kindergarten through grade 12. Reading skill items assess these skills and understanding within the five blueprint domains of the test.

Valid item development is contingent upon several interdependent factors. The following section outlines the factors which guide reading skill item

content development. Item content is comprised of stems, answer choices, and short passages. Additional, detailed information may be found in the English Language Arts Content Appropriateness Guidelines and Item Development Guidelines outlined in the content specification.

Adherence to Skills

Star Reading Spanish assesses more than 600 grade-specific skills within the Renaissance Reading Learning Progression. Item development is skill-specific. Each item in the item bank is developed for and clearly aligned to one skill. An item meets the alignment criteria if the knowledge and skill required to correctly answer the item match the intended knowledge and skill being assessed. Answering an item correctly does not require reading skill knowledge beyond the expected knowledge for the skill being assessed. Star Reading Spanish items include only the information and text needed to assess the skill. All items adhere to skills in ways that are appropriate to the particularities of Spanish. For example, items assessing understanding of idioms present idioms that are specific to the Spanish language.

Level of Difficulty: Readability

Readability is a primary consideration for level of item difficulty. Readability relates to the overall ease of reading an item and involves the reading level, as well as the layout and visual impact of the stem, passage/support information/graphics, and the answer choices. Readability in Star item development accounts for the combined impact, including intensity and density, of each part of the item, even though the individual components of the item may have different readability guidelines.

The reading level and grade level of items is determined by a combination of reference to the Spanish Graded Vocabulary List, editorial judgment of text complexity, and expert review of items for grade-level assignment. Item stems and answer choices present several challenges to accurately determining reading level. Items may contain discipline-specific vocabulary that is typically above grade level but may still be appropriate for the item. Examples of this could include summary, paragraph, or organized and the like. Answer choices may be incomplete sentences for which it is difficult to get an accurate reading of grade level. These factors are taken into account when determining reading level.

Item stems and answer choices that are complete sentences are written for the intended grade level of the item. The words in answer choices and stems that are not complete sentences are within the designated grade-level

range. Reading comprehension is not complicated by unnecessarily difficult sentence structure and/or vocabulary.

Level of Difficulty: Cognitive Load, Content Differentiation, and Presentation

In addition to readability, each item is constructed with consideration to cognitive load, content differentiation, and presentation as appropriate for the ability and experience of a typical student at that grade level.

- **Cognitive Load:** Cognitive load involves the type and amount of knowledge and thinking that a student must have and use in order to answer the item correctly. The entire impact of the stem and answer choices must be taken into account.
- **Content Differentiation:** Content differentiation involves the level of detail that a student must address to correctly answer the item. Determining and/or selecting the correct answer should not be dependent on noticing subtle differences in the stem or answer choices.
- **Presentation:** The presentation of the item includes consistent placement of item components, including directions, stimulus components, questions, and answer choices. Each of these should have a typical representation for the discipline area and grade level. The level of visual differentiation needed to read and understand the item components must be grade-level appropriate.

Efficiency in Use of Student Time

Efficiency is evidenced by a good return of information in relation to the amount of time the student spends on the item. The action(s) required of the student are clearly evident. Ideally, the student is able to answer the question without reading the answer choices. Star Reading Spanish items have clear, concise, precise, and straightforward wording.

Balanced Items: Bias and Fairness

Item development meets established demographic and contextual goals that are monitored during development to ensure the item bank is demographically and contextually balanced. Goals are established and tracked in the following areas: use of fiction and nonfiction text, subject and topic areas, geographic region, gender, ethnicity, occupation, age, and disability.

- Items are free of stereotyping, representing different groups of people in non-stereotypical settings.

- Items do not refer to inappropriate content that includes, but is not limited to content that presents stereotypes based on ethnicity, gender, culture, economic class, or religion.
- Items do not present any ethnicity, gender, culture, economic class, or religion unfavorably.
- Items do not introduce inappropriate information, settings, or situations.
- Items do not reference illegal activities, sinister or depressing subjects, religious activities or holidays based on religious activities, witchcraft, or unsafe activities.

Accuracy of Content

Concepts and information presented in items are accurate, up-to-date, and verifiable. This includes, but is not limited to, references, dates, events, and locations.

Language Conventions

Grammar, usage, mechanics, and spelling conventions in all Star Reading Spanish items adhere to the rules and guidelines in the approved content reference books. *The Dictionary of Spanish Usage* by María Moliner and the *Royal Spanish Academy Dictionary of Spanish Language* are the references for pronunciation, spelling, grammar, mechanics, and usage.

Item Components

In addition to the guidelines outlined above, there are criteria that apply to individual item components. The guidelines for passages are addressed above. Specific considerations regarding stem and distractors are listed below.

Item stems meet the following criteria with limited exceptions:

- The question is concise, direct, and a complete sentence. The question is written so students can answer it without reading the distractors.
- Generally, completion (blank) stems are not used. If a completion stem is necessary, (such as is the case with vocabulary-in-context skills) the stem contains enough information for the student to complete the stem without reading the distractors, and the completion blank is as close to the end of the stem as possible.

- The stem does not include verbal or other clues that hint at correct or incorrect distractors.
- The syntax and grammar are straightforward and appropriate for the grade level. Negative construction is avoided.
- The stem does not contain more than one question or part.
- Concepts and information presented in the items are accurate, up-to-date, and verifiable. This includes but is not limited to dates, references, locations, and events.

Distractors meet the following criteria with limited exceptions:

- All distractors are plausible and reasonable.
- Distractors do not contain clues that hint at correct or incorrect distractors. Incorrect answers are created based on common student mistakes.
- Distractors are independent of each other, are approximately the same length, have grammatically parallel structure, and are grammatically consistent with the stem.
- *Ninguno de estos, ninguno de las respuestas, no se da, todos los respuestas,* y *todos estos* are not used as distractors.

Item and Scale Calibration

Background

Star Reading Spanish, Version 1, was published in 2012 as a measure of reading comprehension. Its item bank contained over 1,800 vocabulary-in-context items, which were calibrated on a vertical scale of difficulty using the Rasch 1-parameter logistic item response model. The current version of Star Reading Spanish, published in 2018, is a broader range, standards-based test; in addition to reading comprehension, its item bank includes questions aligned to four additional content domains. Like its predecessor version, the current Star Reading Spanish uses the calibrated Rasch difficulty of the test items as the basis for adaptive item selection. And it uses the Rasch difficulty of the items administered to a student, along with the pattern of right and wrong answers, to calculate a maximum likelihood estimate of the location of the student on the Rasch scale.

This chapter presents the technical details of the development of the current Star Reading Spanish Rasch scale. Details of the development of the scale for reporting Star Reading Spanish test scores—the Unified Score Scale—will also be presented.

Calibration of Star Reading Spanish Items for Use in Version 2

In Star Reading Spanish Version 2 development, a large-scale item calibration program was completed in the spring of 2018. The Star Reading Spanish 2 item calibration study incorporated all of the newly written and transadapted standards-based Spanish items, as well as over 850 vocabulary items from the Star Reading Spanish 1 item bank that were also recalibrated for use in Star Reading Spanish 2. Two distinct phases comprised the item calibration study. The first phase was the collection of item response data from a multi-level national student sample. The second phase involved the fitting of item response models to the data and developing a single IRT difficulty scale spanning all levels from grades K–12.

Sample Description

The data collection phase of the Star Reading Spanish 2 calibration study began in the fall of 2014 with a total item pool of over 6,200 items. Of

these, 5,166 items were grades K–8 items. Because of the complexity of the transadaptation of the Star Reading English items in Spanish, the calibration study was multifaceted with multiple calibrations conducted. The final calibration study for the viable grade K–8 items was concluded in the spring of 2018. A nationally representative sample of largely Hispanic students tested these items. A total of 121,026 students from 3,948 schools participated in the item calibration study. Table 6 provides the numbers of students in each grade who participated in the study.

Table 6: Numbers of Students Tested by Grade, Star Reading Spanish 2 Item Calibration Study—Spring 2018

Grade Level	Number of Students Tested	Grade Level	Number of Students Tested
1	14,769	5	17,218
2	24,995	6	7,194
3	24,939	7	5,223
4	21,003	8	5,685

The demographics of the calibration sample are summarized in Table 7 and Table 8 below. Table 7 shows that calibration sample comprised primarily of the Hispanic student population. Table 8 shows an almost equal gender split in the calibration sample as expected.

Table 7: Summary of the Calibration Sample by Ethnicity

Ethnicity[a]	Calibration Study
American Indian	4.2%
Asian	1.0%
Black	1.5%
Hispanic	83.2%
White	10.1%

a. There were 51,936 students with no ethnicity reported.

Table 8: Summary of the Calibration Sample by Gender

Gender[a]	Calibration Study
Female	50.3%
Male	49.7%

a. There were 17,326 students with no gender reported.

Item Presentation

For the original calibration research study in 2014, items were tagged with a grade level. The items were then grouped into forms according to grade level while ensuring that each form contained an adequate balance of content measured by Star Reading Spanish. To facilitate vertical scaling, common items (anchors) were included both within grade across the forms (horizontal anchors) and across grades (vertical anchors). The horizontal anchors were used to link forms within grade and the vertical anchors were used to link forms across grade. The vertical anchors were administered at the assigned grade level and one grade level above. The use of anchor items facilitated equating of both test forms and test levels for purposes of data analysis and the development of the overall score scale

Table 9 breaks down the composition of test forms at each grade level in terms of number of test questions, as well as the number of calibration test forms at each level. Students answered a set number of questions at their current grade level, as well as a number of questions one grade level below their grade level.

Table 9: Calibration Test Forms Design by Grade Level, Star Reading Spanish 2 Calibration Study—Spring 2018

Grade Level	Items per Form	Number of Forms
K	30	21
1	35	29
2	35	32
3	35	35
4	45	25
5	45	26
6	45	19
7	45	14
8	45	11
9	45	6
10	45	7
11	45	4
12	45	4
	Sum	233
	× Counterbalancing factor	2
	Total number of forms	466

To avoid problems with positioning effects resulting from the placement of items within each test booklet form, items were shuffled within each test form. This created two variations of each test form such that items appeared in different sequential positions within each "shuffled" test form as indicated by the counterbalancing factor in Table 9. Since the final items would be administered as part of a computer-adaptive test, it was important to remove any effects of item positioning from the calibration data so that each item could be administered at any point during the test.

Calibration test forms were spiraled within the Renaissance calibration software by grade level such that each student received a test form essentially at random. This design ensured that no more than two or three students in any classroom attempted any particular tryout item.

Following that initial calibration of items in grades K–12, items were culled and revised, and a final 2,295 items for grades K–8 underwent the final calibration in the spring of 2018.

Following extensive quality control checks, the Star Reading Spanish 2 calibration research item response data for grades K–8 were analyzed using both traditional item analysis techniques and IRT methods. For each test item, the following information was derived using traditional psychometric item analysis techniques:

- The number of students who attempted to answer the item
- The number of students who did not attempt to answer the item
- The percentage of students who answered the item correctly (a traditional measure of difficulty)
- The percentage of students who selected each answer choice
- The correlation between answering the item correctly and the total score (a traditional measure of item discrimination)
- The correlation between the endorsement of an alternative answer and the total score

Item Difficulty

The difficulty of an item, in traditional item analysis, is the percentage of students who answer the item correctly. This is typically referred to as the "p-value" of the item. Low p-values (such as 15 percent) indicate that the item is difficult since only a small percentage of students answered it correctly. High p-values (such as 90 percent) indicate that almost all students answered the item correctly, and thus the item is easy. It should be noted that the

p-value only has meaning for a particular item relative to the characteristics of the sample of students who responded to it.

Item Discrimination

The traditional measure of the discrimination of an item is the correlation between the "score" on the item (correct or incorrect) and the total test score. Items that correlate well with total test score also tend to correlate well with one another and produce a test that has more reliable scores (more internally consistent). For the correct answer, the higher the correlation between item score and total score, the better the item is at discriminating between low scoring and high scoring students. Such items generally will produce optimal test performance. When the correlation between the correct answer and total test score is low (or negative), it typically indicates that the item is not performing as intended. The correlation between endorsing incorrect answers and total score should generally be low since there should not be a positive relationship between selecting an incorrect answer and scoring higher on the overall test.

Item Response Function

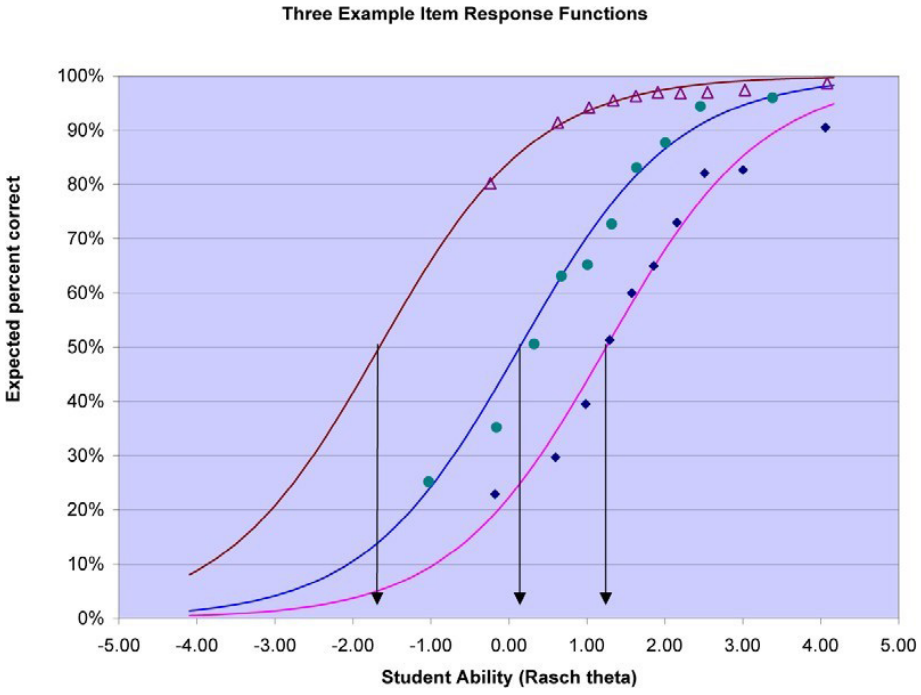
In addition to traditional item analyses, the Star Reading Spanish calibration data were analyzed using Item Response Theory (IRT) methods. Although IRT encompasses a family of mathematical models, the one-parameter (or Rasch) IRT model was selected for the Star Reading Spanish 2 data both for its simplicity and for its ability to accurately model the performance of the Star Reading Spanish 2 items.

IRT attempts to model quantitatively what happens when a student with a specific level of ability attempts to answer a specific question. IRT calibration places the item difficulty and student ability on the same scale; the relationship between them can be represented graphically in the form of an item response function (IRF), which describes the probability of answering an item correctly as a function of the student's ability and the difficulty of the item.

Figure 1 is a plot of three item response functions: one for an easy item, one for a more difficult one, and one for a very difficult item. Each plot is a continuous S-shaped (ogive) curve. The horizontal axis is the scale of student ability, ranging from very low ability (–5.0 on the scale) to very high ability (+5.0 on the scale). The vertical axis is the percent of students expected to answer each of the three items correctly at any given point on the ability scale. Notice that the expected percent correct increases as student ability increases, but varies from one item to another.

In Figure 1, each item's difficulty is the scale point where the expected percent correct is exactly 50. These points are depicted by vertical lines going from the 50 percent point to the corresponding locations on the ability scale. The easiest item has a difficulty scale value of about –1.67; this means that students located at –1.67 on the ability scale have a 50-50 chance of answering that item right. The scale values of the other two items are approximately +0.20 and +1.25, respectively. Calibration of test items estimates the IRT difficulty parameter for each test item and places all of the item parameters onto a common scale. The difficulty parameter for each item is estimated, along with measures to indicate how well the item conforms to (or "fits") the theoretical expectations of the presumed IRT model. Also plotted in Figure 1 are "empirical item response functions (EIRF)": the actual percentages of correct responses of groups of students to all three items. Each group is represented as a small triangle, circle, or diamond. Each of those geometric symbols is a plot of the percent correct against the average ability level of the group. Ten groups' data are plotted for each item; the triangular points represent the groups responding to the easiest item. The circles and diamonds, respectively, represent the groups responding to the moderate and to the most difficult item.

Figure 1: Example of Item Statistics Database Presentation of Information



For purposes of the Star Reading Spanish 2 calibration research, two different "fit" measures (both unweighted and weighted) were computed. Additionally,

if the IRT model is functioning well, then the EIRF points should approximate the (estimated) theoretical IRF. Thus, in addition to the traditional item analysis information, the following IRT-related information was determined for each item administered during the calibration research analyses:

- The IRT item difficulty parameter
- The unweighted measure of fit to the IRT model
- The weighted measure of fit to the IRT model
- The theoretical and empirical IRF plots

Rules for Item Retention

Following these analyses, each test item, along with both traditional and IRT analysis information (including IRF and EIRF plots) and information about the test level, form, and item identifier, were stored in an item statistics database. A panel of content reviewers then examined each item, within content strands, to determine whether the item met all criteria for inclusion into the bank of items that would be used in the norming version of the Star Reading Spanish 2 test.

The item statistics database allowed experts easy access to all available information about an item in order to interactively designate items that, in their opinion, did not meet acceptable standards for inclusion in the Star Reading Spanish 2 item bank. Items were eliminated when they met one or more of the following criteria:

- Item-total correlation (item discrimination) was < 0.30
- Some other answer option had an item discrimination that was high
- Sample size of students attempting the item was less than 200
- The traditional item difficulty indicated that the item was too difficult or too easy
- The item did not appear to fit the Rasch IRT model

For Star Reading Spanish version 2, after each content reviewer had designated certain items for elimination, their recommendations were combined and a second review was conducted to resolve issues where there was not uniform agreement among all reviewers.

Of the initial 5,100+ items tagged with grade levels K–8, only 2,270 remained to be administered in the Star Reading Spanish 2 calibration research study. The item attrition was due to an iterative transadaptation and process

that reviewed both the suitability of content and preliminary psychometric properties of items for operational use in Star Spanish. Of those, 1,498 were deemed of sufficient quality to be retained for operational use.

Scale Calibration and Linking

The outcome of the item calibration study described above was a sizable bank of test items suitable for use in the Star Reading Spanish 2 test, with an IRT difficulty scale parameter for each item. The item difficulty scale itself was devised such that it spanned a range of item difficulty from grades K–8. An important feature of Item Response Theory is that the same scale used to characterize the difficulty of the test items is also used to characterize examinees' ability; in fact, IRT models express the probability of a correct response as a function of the difference between the scale values of an item's difficulty and an examinee's ability. The IRT ability/difficulty scale is continuous; values of observed Rasch ability ranged from about –7.0 to +7.0, with the zero-value occurring at about the third-grade level.

Because of the relationship between Star Reading Spanish and its counterpart Star Reading English, a decision was made to place both tests on a common scale that can be used to report scores on both tests. Such a scale, the Unified Score Scale, has been developed, and was introduced into use in the 2017–2018 school year as the default scale for reporting achievement on Star Reading Spanish tests.

The Unified Score Scale is derived from the Star Reading English Rasch scale of ability and difficulty, which was first introduced with the development of Star Reading English Version 2.

The Star Reading Spanish unified score scale was developed by performing the following steps:

- The Rasch scale used by Star Spanish was linked (transformed) to the Star Reading English Rasch scale.
- A linear transformation of the transformed Rasch scale was developed that spans the entire range of knowledge and skills measured by both Star Reading Spanish and Star Reading English.

Details of these two steps are presented below.

1. The Rasch scale used by Star Reading Spanish was linked to the Star Reading English Rasch scale. In this step, a linear transformation of the Star Reading Spanish Rasch scale to the Rasch scale used by Star Reading English was developed, using a method for linear equating of

IRT (item response theory) scales described by Kolen and Brennan (2004, pages 161–165). The linear equating process used all of the common items between Star Reading Spanish and Star Reading English. Because Renaissance calibrates items and persons on the same scale using the Rasch model, the linking equation developed based on common items could be used to transform the student scores from the Spanish scale to the English scale.

2. Because Rasch scores are expressed as decimals, and may be either negative or positive, a more user-friendly scale score was developed that uses positive integer numbers only. A linear transformation of the extended Star Reading Spanish Rasch scale was developed that spans the entire range of knowledge and skills measured by both Star Reading Spanish and Star Reading English. The transformation formula is as follows:

$$\text{Unified Scale Score} = \text{INT}(42.93 * \text{Star Reading Spanish Rasch Score} + 958.74)$$

Reported Star Reading Spanish unified scale scores range from 600–1400.

On-line Data Collection for New Item Calibration

Beginning with the 2018–2019 school year, new test items at grade levels K–12 are being developed and calibrated for use in Star Reading Spanish. The data needed for item calibration are collected on-line, by embedding small numbers of uncalibrated items within Star Reading Spanish tests. After sufficient numbers of item responses have accumulated, the Rasch difficulty of each new item is estimated by fitting a logistic model to the item response data and the Star Reading Spanish Rasch scores of the student's tests. Renaissance calls this overall process "dynamic calibration."

Typically, dynamic calibration is done in batches of several hundred new test items. Each student's test may include between 1 and 3 uncalibrated items.

Each item is tagged with a grade level, and is typically administered only to students at that grade level and the next higher grade. The selection of the uncalibrated items to be administered to each student is at random, resulting in nearly equivalent distributions of student ability for each item at a given grade level.

Both traditional and IRT item analyses are conducted of the item response data collected. The traditional analyses yield proportion correct statistics,

as well as biserial and point-biserial correlations between scores on the new items and actual scores on the Star Reading Spanish tests.

For dynamic calibration, a minimum of 1,000 responses per item is the data collection target. In practice, because of the number of Star Reading Spanish tests administered each year, the number of students responding to each new test item is expected to equal or exceed the target. The calibration analysis proceeds one item at a time, using SAS/STAT™ software to estimate the threshold (difficulty) parameter of every new item by calculating the non-linear regression of each new item score (0 or 1) on the Star Reading Spanish Rasch ability estimates. The accuracy of the non-linear regression approach has been corroborated by conducting parallel analyses using Winsteps software. In tests, the two methods yielded virtually identical results.

Computer-Adaptive Test Design

In computer-adaptive tests like the Star Reading Spanish test, the items taken by a student are dynamically selected in light of that student's performance during the testing session. Thus, low-performing students may be administered easier items in order to better estimate their reading achievement level. High-performing students may branch to more challenging reading items in order to better determine the breadth of their reading skills and their reading achievement level.

During a Star Reading Spanish test, a student may be "routed" to items at the lowest reading level or to items at higher reading levels within the overall pool of items, depending on the student's unfolding performance during the testing session. In general, when an item is answered correctly, the student is then given a more difficult item. When an item is answered incorrectly, the student is then given an easier item. Item difficulty here is defined by results of the Star Reading Spanish item calibration studies.

Students who have not taken a Star Reading Spanish test within six months initially receive an item whose difficulty level is relatively easy for students at the examinee's grade level. The selection of an item that is a bit easier than average minimizes any effects of initial anxiety that students may have when starting the test and serves to better facilitate the student's initial reactions to the test. These starting points vary by grade level and were based on research conducted as part of the national item calibration study.

When a student has taken a Star Reading Spanish test within the last six months, the difficulty of the first item depends on that student's previous Star Reading Spanish test score information. After the administration of the initial

item, and after the student has entered an answer, Star Reading Spanish software estimates the student's reading ability. The software then selects the next item randomly from among all of the items available that closely match the student's estimated reading ability.

Random selection from among items with difficulty values near the student's adjusted reading ability allows the program to avoid overexposure of test items. Items that have been administered to the same student within the past three-month time period are not available for administration. The large numbers of items available in the item pools, however, ensure that this constraint has negligible impact on the quality of each Star Reading Spanish computer-adaptive test.

Scoring in the Star Reading Spanish Tests

Following the administration of each Star Reading Spanish item, and after the student has selected an answer, an updated estimate of the student's reading ability is computed based on the student's responses to all items that have been administered up to that point. A proprietary Bayesian-modal Item Response Theory (IRT) estimation method is used for scoring until the student has answered at least one item correctly and one item incorrectly. Once the student has met the 1-correct/1-incorrect criterion, Star Reading Spanish software uses a proprietary Maximum-Likelihood IRT estimation procedure to avoid any potential of bias in the Scaled Scores.

This approach to scoring enables Star Reading Spanish to provide Scaled Scores that are statistically consistent and efficient. Accompanying each Scaled Score is an associated measure of the degree of uncertainty, called the conditional standard error of measurement (CSEM). Unlike a conventional paper-and-pencil test, the CSEM values for the Star Reading Spanish test are unique for each student. CSEM values are dependent on the particular items the student received and on the student's performance on those items.

Scaled Scores are expressed on a common scale that spans all grade levels covered by Star Reading Spanish (grades K–8). Because of this common scale, Scaled Scores are directly comparable with each other, regardless of grade level. Other scores, such as Percentile Ranks and Grade Equivalents, are derived from the Scaled Scores.

Reliability and Measurement Precision

Measurement is subject to error. A measurement that is subject to a great deal of error is said to be *imprecise*; a measurement that is subject to relatively little error is said to be *reliable*. In psychometrics, the term *reliability* refers to the degree of measurement precision, expressed as a proportion. A test with perfect score precision would have a reliability coefficient equal to 1, meaning that 100 percent of the variation among persons' scores is attributable to variation in the attribute the test measures, and none of the variation is attributable to error. Perfect reliability is probably unattainable in educational measurement; for example, a test with a reliability coefficient of 0.90 is more likely. On such a test, 90 percent of the variation among students' scores is attributable to the attribute being measured, and 10 percent is attributable to errors of measurement. Another way to think of score reliability is as a measure of the consistency of test scores. Two kinds of consistency are of concern when evaluating a test's measurement precision: internal consistency and consistency between different measurements. First, internal consistency refers to the degree of confidence one can have in the precision of scores from a single measurement. If the test's internal consistency is 95 percent, just 5 percent of the variation of test scores is attributable to measurement error.

Second, reliability as a measure of consistency between two different measurements indicates the extent to which a test yields consistent results from one administration to another and from one test form to another. Tests must yield somewhat consistent results in order to be useful; the reliability coefficient is obtained by calculating the coefficient of correlation between students' scores on two different occasions, or on two alternate versions of the test given at the same occasion. Because the amount of the attribute being measured may change over time, and the content of tests may differ from one version to another, the internal consistency reliability coefficient is generally higher than the correlation between scores obtained on different administrations.

There are a variety of methods of estimating the reliability coefficient of a test. Methods such as Cronbach's alpha and split-half reliability are single administration methods and assess internal consistency. Coefficients of correlation calculated between scores on alternate forms, or on similar tests administered two or more times on different occasions, are used to assess alternate forms reliability, or test-retest reliability (stability).

In a computerized adaptive test such as Star Reading Spanish, content varies from one administration to another, and it also varies with each student's performance.

Another feature of computerized adaptive tests based on Item Response Theory (IRT) is that the degree of measurement error can be expressed for each student's test individually.

The Star Reading Spanish tests provide two ways to evaluate the reliability of scores: reliability coefficients, which indicate the overall precision of a set of test scores, and standard errors of measurement (SEM), which provide an index of the degree of error in test scores.

A reliability coefficient is a summary statistic that reflects the average amount of measurement precision in a specific examinee group or in a population as a whole.

In Star Reading Spanish, two types of SEM are calculated: "global SEM," which is a summary of a test's measurement error, calculated for a sample or population of examinees; and "conditional SEM," CSEM. CSEM is an estimate of the measurement error in each individual test score. While a reliability coefficient is a single value that applies to the test in general, the magnitude of the CSEM may vary substantially from one person's test score to another's.

This chapter presents three different types of reliability coefficients: generic reliability, split-half reliability, and alternate forms (test-retest) reliability. This is followed by statistics on the conditional standard error of measurement and the global standard error of measurement of Star Reading Spanish test scores.

Generic Reliability

Test reliability is generally defined as the proportion of test score variance that is attributable to true variation in the trait the test measures. This can be expressed analytically as

$$Reliability = 1 - \frac{\sigma^2_{error}}{\sigma^2_{total}}$$

where σ^2_{error} is the variance of the errors of measurement, and σ^2_{total} is the variance of test scores. In Star Reading Spanish, the variance of the test scores is easily calculated from Scaled Score data. The variance of the errors of measurement may be estimated from the conditional standard error of

measurement (CSEM) statistics that accompany each of the IRT-based test scores, including the Scaled Scores, as depicted below.

$$\sigma^2_{error} = \frac{1}{n}\sum_{i=1}^{n} SEM_i^2$$

where the summation is over the squared values of the reported CSEM for students i = 1 to n. In each Star Reading Spanish test, CSEM is calculated along with the IRT ability estimate and Scaled Score. Squaring and summing the CSEM values yields an estimate of total squared error; dividing by the number of observations yields an estimate of mean squared error, which in this case is tantamount to error variance. "Generic" reliability is then estimated by calculating the ratio of error variance to Scaled Score variance, and subtracting that ratio from 1.

Using this technique with a stratified random sample of Star Reading Spanish 2016–2017 and 2017–2018 school year data resulted in the generic reliability estimates shown in Table 10. Results in Table 10 indicate that the overall reliability of the unified scale scores was 0.97. Coefficients ranged from a low of 0.91 in grade 1 to a high of 0.96 in grades 7 and 8. Because this method is not susceptible to error variance introduced by repeated testing, multiple occasions, and alternate forms, the resulting estimates of reliability are generally higher than the more conservative alternate forms reliability coefficients. These generic reliability coefficients are, therefore, plausible upper-bound estimates of the internal consistency reliability of the Star Reading Spanish computer-adaptive test.

Table 10: Reliability Estimates from the Star Reading Spanish 2016–2017 and 2017–2018 Data on the Unified Scale

	Reliability Estimates: For Unified Scale						
	Generic		Split-Half		Alternate Forms		
Grade	N	ρxx	N	ρxx	N	ρxx	Average Days between Testing
1	16,000	0.91	16,000	0.87	1,705	0.69	88
2	16,000	0.94	16,000	0.91	1,705	0.76	89
3	16,000	0.95	16,000	0.93	1,705	0.78	92
4	16,000	0.95	16,000	0.93	1,705	0.82	89
5	16,000	0.95	16,000	0.94	1,705	0.82	90
6	16,000	0.95	16,000	0.94	1,705	0.80	95
7	8,000	0.96	8,000	0.94	1,405	0.80	113
8	8,000	0.96	8,000	0.94	1,405	0.81	107
Overall	112,000	0.97	112,000	0.95	13,040	0.88	95

As the data in Table 10 shows, Star Reading Spanish reliability is high, grade by grade and overall. Star Reading Spanish's technical quality for an interim assessment is on a virtually equal footing with the highest-quality summative assessments in use today.

Split-Half Reliability

While generic reliability does provide a plausible estimate of measurement precision, it is a theoretical estimate, as opposed to traditional reliability coefficients, which are more firmly based on item response data. Traditional internal consistency reliability coefficients such as Cronbach's alpha and Kuder-Richardson Formula 20 (KR-20) are not meaningful for adaptive tests. However, an estimate of internal consistency reliability can be calculated using the split-half method.

A split-half reliability coefficient is calculated in three steps. First, the test is divided into two halves, and scores are calculated for each half. Second, the correlation between the two resulting sets of scores is calculated; this correlation is an estimate of the reliability of a half-length test. Third, the resulting reliability value is adjusted, using the Spearman-Brown formula,[1] to estimate the reliability of the full-length test.

In internal simulation studies, the split-half method provided accurate estimates of the internal consistency reliability of adaptive tests, and so it has been used to provide estimates of Star Reading Spanish reliability. These split-half reliability coefficients are independent of the generic reliability approach discussed earlier and more firmly grounded in the item response data.

Split-half scores were based on all of the 34 items of the Star Reading Spanish tests; scores based on the odd- and the even-numbered items were calculated separately. The correlations between the two sets of scores were corrected to a length of 34 items, yielding the split-half reliability estimates displayed in Table 10.

Results indicated that the overall split-half reliability of the Unified scores was 0.95. The coefficients ranged from a low of 0.87 in grade 1 to a high of 0.94 in grades 5 to 8. These reliability estimates are quite consistent across grades 1–8, and quite high, again a result of the measurement efficiency inherent in the adaptive nature of the Star Reading Spanish test.

1. See Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley, pp. 112–113.

Alternate Forms Reliability

Another method of evaluating the reliability of a test is to administer the test twice to the same examinees. Next, a reliability coefficient is obtained by calculating the correlation between the two sets of test scores. This is called a test-retest reliability coefficient if the same test was administered both times and an alternate forms reliability coefficient if different, but parallel, tests were used.

Content sampling, temporal changes in individuals' performance, and growth or decline over time can affect alternate forms reliability coefficients, usually making them appreciably lower than internal consistency reliability coefficients.

The alternate forms reliability study provided estimates of Star Reading Spanish reliability using a variation of the test-retest method. In the traditional approach to test-retest reliability, students take the same test twice, with a short time interval, usually a few days, between administrations. In contrast, the Star Reading Spanish alternate form reliability study administered two different tests by avoiding during the second test the use of any items the student had encountered in the first test. All other aspects of the two tests were identical. The correlation coefficient between the scores on the two tests was taken as the reliability estimate.

The alternate forms reliability estimates for the Star Reading Spanish test were calculated using the Star Reading Spanish Unified scaled scores. Checks were made for valid test data on both test administrations and to remove cases of apparent motivational discrepancies.

Table 10 includes overall and within-grade alternate reliability, along with an indication of the average number of days between testing occasions. The average number of days between testing occasions ranged from 85–113 days. Results indicated that the overall reliability of the scores on the Unified scale was about 0.88. The alternate forms coefficients ranged from a low of 0.69 in grade 1 to a high of 0.82 in grades 4 and 5.

Because errors of measurement due to content sampling and temporal changes in individuals' performance can affect this correlation coefficient, this type of reliability estimate provides a conservative estimate of the reliability of a single Star Reading Spanish administration. In other words, the actual Star Reading Spanish reliability is likely higher than the alternate forms reliability estimates indicate.

Standard Error of Measurement

When interpreting the results of any test instrument, it is important to remember that the scores represent estimates of a student's true ability level. Test scores are not absolute or exact measures of performance. Nor is a single test score infallible in the information that it provides. The standard error of measurement (SEM) can be thought of as a measure of how precise a given score is; smaller values of SEM or CSEM indicate greater precision.

The standard error of measurement describes the extent to which scores would be expected to fluctuate because of chance. If measurement errors follow a normal distribution, an SEM of 18 means that if a student were tested repeatedly, his or her scores would fluctuate within 18 points of his or her first score about 68 percent of the time, and within 36 points (twice the SEM) roughly 95 percent of the time. Since reliability can also be regarded as a measure of precision, there is an inverse relationship between the reliability of a test and the standard error of measurement for the scores it produces: lower standard error of measurement results in higher reliability.

The Star Reading Spanish tests differ from traditional tests in at least two respects with regard to the standard error of measurement. First, Star Reading Spanish software computes the SEM for each individual student based on his or her performance, unlike most traditional tests that report the same SEM value for every examinee. Each administration of Star Reading Spanish yields a unique "conditional" SEM (CSEM) that reflects the amount of information estimated to be in the specific combination of items that a student received in his or her individual test. Second, because the Star Reading Spanish test is adaptive, the CSEM will tend to be lower than that of a conventional test of the same length, particularly at the highest and lowest score levels, where conventional tests' measurement precision is weakest. Because the adaptive testing process attempts to provide equally precise measurement, regardless of the student's ability level, the average CSEMs for the IRT ability estimates are generally similar for all students.

Table 11 contains two different sets of estimates of Star Reading Spanish measurement error: conditional standard error of measurement (CSEM) and global standard error of measurement (SEM). Conditional SEM was just described; the estimates of CSEM in Table 11 are the average CSEM values observed for each grade.

Global standard error of measurement is based on the traditional SEM estimation method, using the estimated generic reliability and the variance of the test scores to estimate the SEM:

$$SEM = \sqrt{1 - \rho_{xx}}\sigma_x$$

where, ρ_{xx} is the estimated generic reliability, and σ_x is the standard deviation of the observed scores (in this case, Scaled Scores).

Table 11 summarizes the distribution of CSEM values for the 2016–2017 and 2017–2018 data, overall and by grade level. The overall average CSEM on the Unified scale across all grades was 18 scaled score units as was the average CSEM in grades 1–8.

Table 11 also shows the estimates of the global SEM. The global SEM estimates were slightly higher than the CSEM estimates. The overall average was 20, while the SEM estimates were 20 for grades 2–8 and 21 for grade 1.

Table 11: Standard Error of Measurement for the 2016–2017 and 2017–2018 Star Reading Spanish data on the Unified Scale

	Standard Error of Measurement Unified Scale				
	Conditional			Global	
Grade	N	Average	Standard Deviation	N	SEM
1	16,000	18	1.3	16,000	21
2	16,000	18	1.1	16,000	20
3	16,000	18	1.2	16,000	20
4	16,000	18	1.1	16,000	20
5	16,000	18	1.3	16,000	20
6	16,000	18	1.6	16,000	20
7	8,000	18	1.6	8,000	20
8	8,000	18	2.0	8,000	20
All	112,000	18	1.4	112,000	20

Validity

Test validity was long described as the degree to which a test measures what it is intended to measure. An updated conceptualization of test validity is that test validity consists of the collection of evidentiary data to support specific claims as to *what* the test measures, the *interpretation* of its scores, and the *uses* for which it is recommended or applied. Evidence of test validity is often indirect and incremental, consisting of a variety of data that in the aggregate are consistent with the theory that the test measures the intended construct(s), or is suitable for its intended uses and interpretations of its scores. Determining whether there is test validity evidence to support the intended uses and interpretations of test scores involves the use of data and other information both internal and external to the test instrument itself.

Content Validity

One touchstone is content validity, which is the relevance of the test questions to the attributes or dimensions intended to be measured by the test—namely reading comprehension, reading vocabulary, and related reading skills, in the case of the Star Reading Spanish assessments. The content of the item bank and the content balancing specifications that govern the administration of each test together form the foundation for "content validity" for the Star Reading Spanish assessments. These content topics were discussed in detail in "Content and Item Development" and were an integral part of the test items that are the basis of Star Reading Spanish today.

Construct Validity

Construct validity, which is the overarching criterion for evaluating a test, investigates the extent to which a test measures the construct(s) that it claims to be assessing. Establishing construct validity involves the use of data and other information external to the test instrument itself. For example, Star Reading Spanish claims to provide an estimate of a child's reading comprehension and achievement level. Therefore, demonstration of Star Reading Spanish's construct validity rests on the evidence that the test provides such estimates. There are a number of ways to demonstrate this.

This section deals with both internal and external evidence of the validity of Star Reading Spanish as an assessment of reading comprehension and reading skills.

Internal Evidence

Evaluation of Unidimensionality of Star Reading Spanish

Star Reading Spanish is a 34-item computerized-adaptive assessment that measures reading comprehension. Its items are selected adaptively for each student, from a very large bank of reading test items, each of which is aligned to one of five blueprint domains:

- Word knowledge and skills,
- Comprehension strategies and constructing meaning,
- Analyzing literary text,
- Analyzing argument and evaluating text, and
- Understanding author's craft.

Star Reading Spanish is an application of item response theory (IRT); each test item's difficulty has been calibrated using the Rasch 1-parameter logistic IRT model. One of the assumptions of the Rasch model is unidimensionality: that a test measures only a single construct such as reading comprehension in the case of Star Reading Spanish. To evaluate whether Star Reading Spanish measures a single construct, factor analyses were conducted. Factor analysis is a statistical technique used to determine the number of dimensions or constructs that a test measures. Both exploratory and confirmatory factor analyses were conducted across grades 1 to 8.

To begin, a large sample of student Star Reading Spanish data was assembled. The overall sample consisted of 284,734 student records in the 2016–2017 or 2017–2018 school years. From that sample, stratified random samples of 10,000 students per grade were taken to yield a sample of 80,000 students for analysis. These data were the focus of the exploratory and confirmatory factor analyses.

Prior to performing the factor analyses, each student's 34 Star Reading Spanish item responses were divided into subsets of items aligned to each of the 5 blueprint domains. Tests administered in grades 4–8 included items from all five domains. Tests given in grades 1–3 included items from just 4 domains; no items measuring analyzing argument and evaluating text were administered in these grades.

For each student, separate Rasch ability estimates (subtest scores) were calculated from each domain-specific subset of item responses. A Bayesian sequential procedure developed by Owen (1969, 1975) was used for the subtest scoring. The number of items included in each subtest ranged from 2 to 18,

following the Star Reading Spanish test blueprints, which specify different numbers of items per domain, depending on the student's grade level.

Inter-correlations of the blueprint domain-specific Rasch subtest scores were analyzed using exploratory factor analysis (EFA) to evaluate the number of dimensions/ factors underlying Star Reading Spanish. Varimax rotation was used. The EFA retained a single dominant underlying dimension based on either the MINEIGEN (eigenvalue greater than 1) or the PROPORTION criterion (proportion of variance explained by the factor), as expected.

Figure 2 and Figure 3 show the scree plots and variance explained per factor for the combined analyses of grades 1 through 3 and grades 4 through 8, respectively.

Figure 2: Scree Plot and Variance Explained by Factor Plot from the Grades 1 through 3 Exploratory Factor Analysis in Star Reading Spanish

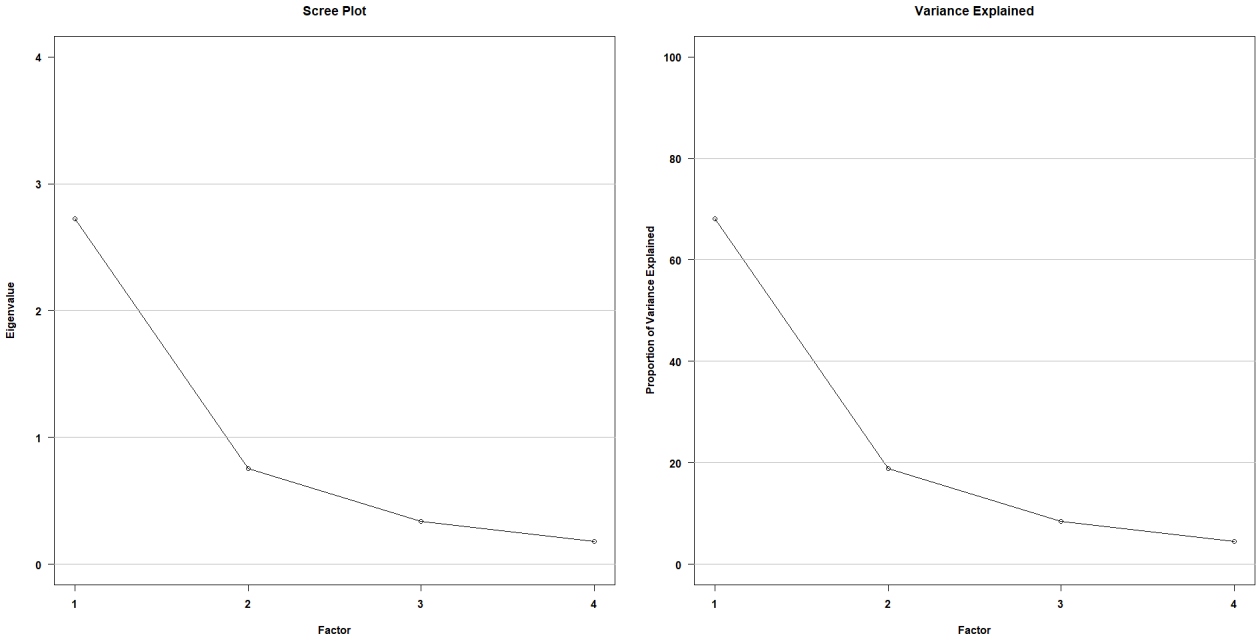
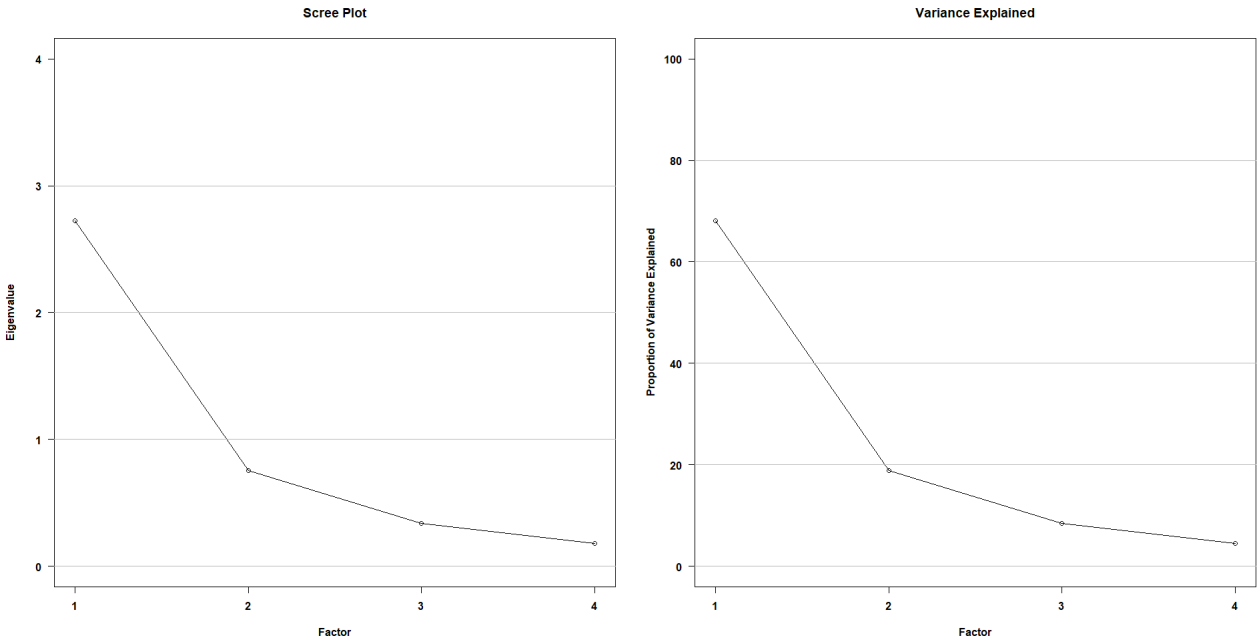
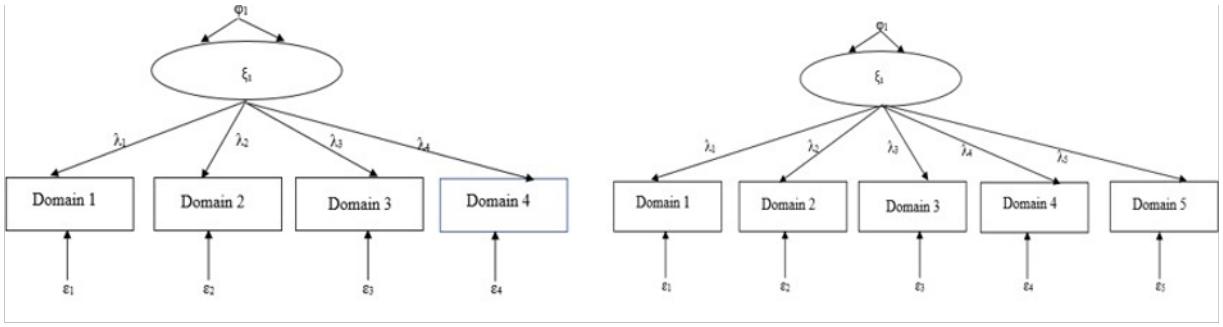


Figure 3: Scree Plot and Variance Explained by Factor Plot from the Grades 4 through 8 Exploratory Factor Analysis in Star Reading Spanish



Subsequent to the EFA, confirmatory factor analyses (CFA) were also conducted using the subtest scores from the CFA sub-sample. A separate confirmatory analysis was conducted for each grade. The CFA models tested a single underlying model as shown in Figure 4. Two CFA models were fitted because one of the Star Reading Spanish blueprint domains is not tested in grades 1 through 3.

Figure 4: Confirmatory Factor Analyses (CFA) in Star Reading Spanish



Model 1: Grades 1 through 3

Model 2: Grades 4 through 8

The results of the CFA are summarized in Table 12 below. As the table indicates, the sample size for each grade was 10,000; because the chi-square (X^2) test is not a reliable test of model fit when sample sizes are large, fit indices are presented. The comparative fit index (CFI) and the Tucker-Lewis index (TLI) are shown; for these indices, values are either 1 or very close to 1, indicating strong evidence of a single construct/dimension. In addition, the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) are presented. RMSEA and SRMR values less than 0.08 indicate good fit. Cutoffs for the indices are presented in Hu and Bentler (1999). Overall, the CFA results strongly support a single underlying construct in Star Reading Spanish.

Table 12: Summary of the Goodness-of-Fit of the CFA Models for Star Reading Spanish by Grade

Grade	N	X^2	df	CFI	TLI	RMSEA	SRMR
1	10,000	17.52	2	1.00	1.00	0.03	0.01
2	10,000	0.08	2	1.00	1.00	0.00	0.00
3	10,000	10.62	2	1.00	1.00	0.02	0.00
4	10,000	83.39	5	1.00	1.00	0.04	0.01
5	10,000	102.59	5	1.00	1.00	0.04	0.01
6	10,000	149.57	5	1.00	0.99	0.05	0.01
7	10,000	188.43	5	1.00	0.99	0.06	0.01
8	10,000	215.34	5	1.00	0.99	0.07	0.01
Grades 1 to 3	30,000	10.61	2	1.00	1.00	0.01	0.00
Grades 4 to 8	50,000	748.57	5	1.00	0.99	0.06	0.01

The EFA were conducted using the factanal function in R 3.5.1 (R Core Team, 2018), while the CFA were conducted using the lavaan package (Rosseel, 2012) in R.

Another part of assessing the dimensionality of the Star Reading Spanish is looking at the measurement invariance of the assessments across grades that share the same blueprint. There are four different models of measurement invariance that can be tested to see whether they hold across grades. The most constrained model is called strict measurement invariance, where the factor loadings, intercepts, and residuals are constrained to be equal across grades. If only the factor loadings and intercepts are constrained to be equal, it is called strong measurement invariance, and if only the loadings are constrained to be equal, it is called weak measurement invariance. Configural measurement invariance is the weakest type of

measurement invariance, where there is the same pattern of loadings across grades, but there are no equality constraints. Given that the Star Reading Spanish assessment is fit with the Rasch model using a single underlying vertical scale and the levels of performance across grades sometimes differ, the configural and weak measurement invariance models should hold, but the strong and strict measurement invariance models may not hold.

Table 13 shows the measurement invariance models and fit statistics for grades 1–3, and Table 14 shows the measurement invariance models and fit statistics for grades 4–8. The results in the tables suggest that all four types of the measurement invariance models exhibit reasonable fit. These results provide additional support that the construct assessed by the Star Reading Spanish assessments is consistent across grades and that application of the Rasch model and a single vertical scale is appropriate.

Table 13: Measurement Invariance Statistics for Star Reading Spanish for Grades 1–3

Model Type	N	X^2	df	CFI	TLI	RMSEA	SRMR
Configural	30,000	28.22	6	1.00	1.00	0.02	0.00
Weak	30,000	111.19	12	1.00	1.00	0.03	0.02
Strong	30,000	242.41	18	1.00	1.00	0.04	0.02
Strict	30,000	328.97	26	1.00	1.00	0.04	0.02

Table 14: Measurement Invariance Statistics for Star Reading Spanish for Grades 4–8

Model Type	N	X^2	df	CFI	TLI	RMSEA	SRMR
Configural	50,000	739.32	25	1.00	0.99	0.05	0.01
Weak	50,000	1469.93	41	0.99	0.99	0.06	0.03
Strong	50,000	2088.44	57	0.99	0.99	0.06	0.04
Strict	50,000	2222.55	77	0.99	0.99	0.06	0.04

Types of External Evidence

In an ongoing effort to gather evidence for the validity of Star Reading Spanish scores, continual research on score validity has been undertaken. In addition to original validity data gathered at the time of initial development, a small number of studies have investigated correlations between Star Reading Spanish tests and other external measures. There are generally three types of correlations with external measures that can be explored: concurrent validity estimates, predictive validity estimates, and discriminant validity estimates.

For Star Reading Spanish, concurrent validity is defined as taking a Star Reading Spanish test and another external measure that also assesses reading achievement in Spanish within a month time period. At present, only a small number of concurrent validity studies have been conducted since Star Reading Spanish has only been used operationally for a few years. Predictive validity provides estimates of the extent to which scores on the Star Reading Spanish test predict scores on an external measure of reading achievement in Spanish at a later point in time, operationally defined as more than a month between the Star test (predictor) and the criterion test. No studies of the predictive validity of Star Reading Spanish have yet been conducted. Future studies will explore the predictive validity of Star Reading Spanish as the test continues to be used. Discriminant validity estimates consist of taking Star Reading Spanish and another external measure that assess another content area besides reading achievement in Spanish (e.g., correlations with a math achievement measure) within a month time period. Typically, the goal is that discriminant validity estimates are lower than concurrent validity estimates. Only a small number of discriminant validity estimates have been collected.

External Evidence

Relationship of Star Reading Spanish Scores to Other Tests of Spanish Reading Achievement

As of the end of 2018, one study has correlated Star Reading Spanish results with two Spanish reading subtest scores for easyCBM, and another study has correlated Star Reading Spanish results with the State of Texas Assessments of Academic Readiness (STAAR) Reading Spanish test to provide concurrent validity estimates. Table 15 provides a summary of those analyses. The easyCBM study took place during the 2015–2016 school year and the STAAR study took place in the Spring of 2018. Concurrent validity estimates with easyCBM ranged from 0.58 to 0.67 and concurrent validity estimates with STAAR ranged from 0.62 to 0.68. These coefficients provide solid evidence of the external relationship between the Star Reading Spanish assessments and these other two Spanish reading assessments.

Table 15: Correlations Between STAR Reading Spanish and Other Spanish Reading Achievement Measures

Test Form	Date	Score	1		2		3		4		5		6		7		8	
			n	r	n	r	n	r	n	r	n	r	n	r	n	r	n	r
State of Texas Assessments of Academic Readiness Standards Test (STAAR)																		
STAAR Reading Spanish	Spring 2018	SS	–	–	–	–	7,203	0.68	4,985	0.67	84	0.62	–	–	–	–	–	–
easyCBM																		
Spanish word reading	2015–2016	SS	–	–	729	0.63	533	0.58	–	–	–	–	–	–	–	–	–	–
Spanish sentence reading	2015–2016	SS	–	–	728	0.67	532	0.65	–	–	–	–	–	–	–	–	–	–

Relationship of Star Reading Spanish to Other Achievement Tests Measuring Math Achievement

As of the end of 2018, five studies have examined the relationship of Star Reading Spanish with other achievement tests measuring math content to provide discriminant validity estimates. One study looked at the relationship between Star Reading Spanish and the Common Core State Standards Math subtest score with Spanish translations for easyCBM, two studies looked at the relationship between Star Reading Spanish and STAAR math Spanish tests, and two studies looked at the relationship between Star Reading Spanish and Star Math Spanish by correlating results for students' first and last assessments taken. Table 16 provides a summary of those analyses. Discriminant validity estimates with the easyCBM Common Core State Standards Math subtest score ranged from 0.39 to 0.54, discriminant validity estimates with STAAR ranged from 0.51 to 0.59, and discriminant validity estimates with Star Math Spanish ranged from 0.44 to 0.62. These discriminant validity estimates show that the relationship of Star Reading Spanish with achievement tests measuring content other than Spanish Reading achievement in several cases were slightly lower than the concurrent validity estimates with Spanish Reading achievement measures. These coefficients provide some evidence of expected external relationships between Star Reading Spanish assessments and these other achievement tests measuring content other than Spanish Reading achievement.

Table 16: Correlations between Star Reading Spanish and Other Achievement Tests Measuring Content Other than Spanish Reading Achievement

			1		2		3		4		5		6		7		8	
Test Form	**Date**	**Score**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**
easyCBM																		
Common Core State Standard Math Score	2015–2016	SS	–	–	930	0.40	696	0.54	164	0.39	–	–	–	–	–	–	–	–
State of Texas Assessments of Academic Readiness Standards Test (STAAR)																		
STAAR Math Spanish	Spring 2018	SS	–	–	–	–	6,855	0.57	4,680	0.57	81	0.51	–	–	–	–	–	–
STAAR Math English	Spring 2018	SS	–	–	–	–	305	0.59	115	0.57	–	–	–	–	–	–	–	–
Renaissance Star Assessments																		
Star Math Spanish first assessment taken	2016–2017	SS	2,419	0.48	6,868	0.54	5,268	0.54	3,086	0.55	1,650	0.52	570	0.56	499	0.54	438	0.58
Star Math Spanish last assessment taken	2016–2017	SS	1,659	0.45	6,626	0.53	5,669	0.54	3,155	0.56	2,327	0.52	498	0.44	499	0.49	698	0.52
Star Math Spanish first assessment taken	2017–2018	SS	–	–	–	–	3,350	0.51	1,339	0.51	107	0.55	–	–	–	–	–	–
Star Math Spanish last assessment taken	2017–2018	SS	–	–	–	–	3,492	0.62	1,373	0.62	95	0.61	–	–	–	–	–	–

Summary of Star Reading Spanish Validity Evidence

The validity data presented in this technical documentation includes evidence of Star Reading Spanish's content and construct validity. While the amount of data presented in this technical report is less than the amount of data provided for Star Reading since the test has only been in operation for a few years, the data provided is quite positive. The information presented in the "Content and Item Development" chapter supported the content validity of the Star Reading Spanish. Exploratory and confirmatory factor analyses provided evidence that Star Reading Spanish measures a unidimensional construct, consistent with the assumption underlying its use of the Rasch 1-parameter logistic item response model, while measurement invariance analyses provided further evidence to support the use of a single vertical scale and the Rasch model. The small number of concurrent and discriminant validity estimates indicate that Star Reading Spanish exhibits appropriate moderate to high correlations with other measures of Spanish Reading achievement and that these correlations in several cases were slightly higher than correlations with achievement measures in other subjects. Taken together, these data provide support for the claim that Star Reading Spanish is a measure of reading comprehension in Spanish.

Norming

Two distinct kinds of norms are described in this chapter: test score norms and growth norms. The former refers to distributions of test scores themselves. The latter refers to distributions of changes in test scores over time; such changes are generally attributed to growth in the attribute that is measured by a test. Hence distributions of score changes over time may be called "growth norms."

The 2020 Star Reading Spanish Norms

New US norms for Star Reading Spanish assessments were introduced at the start of the 2020–21 school year. Separate early fall and late spring norms were developed for grades 1–8.

The norms introduced in 2020 are based on test scores of grades 1–8 students that took the Star Reading Spanish test during the 2018–2019 school year who had complete assessment data. These norms are on the Star Unified scale.

Students participating in the norming study took assessments between August 1, 2018 and June 30, 2019. Students took the Star Reading Spanish tests under normal test administration conditions. No specific norming test was developed, and no deviations were made from the usual test administration. Thus, students in the norming sample took Star Reading Spanish tests as they are administered in everyday use.

Sample Characteristics

During the norming period, a total of 173,997 US students in grades 1–8 took the Star Reading Spanish tests. The first step in sampling was to select a representative sample of students who had tested in the fall, in the spring, or in both the fall and spring of the 2018–2019 school year under normal testing conditions and who had complete assessment data. Data used for the norming analyses consisted of the full sample of students that took the test in either the fall or the spring. If a student took more than one assessment in the fall, the first assessment administered in the fall was included in the norming sample, and if a student took more than one assessment in the spring, the last assessment taken was included in the norming sample. Since there is not currently a widely accepted definition of what constitutes a representative national population of US students taking Spanish tests, Renaissance's post-

stratification procedure used with Star Reading, Star Early Literacy, and Star Math to make norms nationally representative was not applied to these data. However, data on the percentages in different geographic regions, school enrollments, socioeconomic statuses, school locations, and school types are provided.

The final norming sample size after selecting only students with test scores in the fall, the spring, or both fall and spring in the norming years was 125,605 students in grades 1–8. There were 92,750 students in the fall norming sample and 83,530 students in the spring norming sample. Some students contributed test results in both the fall and spring in the 2018–2019 school year. These students were counted for each unique assessment in each school year when computing the norming sample size. These students came from schools across the 50 US states and the District of Columbia. Table 17 and Table 18 provide a breakdown of the number of students participating per grade in the fall and in the spring, respectively.

Table 17: N Counts per Grade in the Fall Norms Sample

Grade	N
1	6,730
2	22,180
3	21,600
4	17,410
5	13,020
6	5,200
7	3,290
8	3,320
Total	92,750

Table 18: N Counts per Grade in the Spring Norms Sample

Grade	N
1	10,340
2	20,710
3	18,610
4	14,790
5	10,380
6	3,990
7	2,600
8	2,110
Total	83,530

Estimates of US student population characteristics for the schools included in the norming sample were obtained from the Market Data Retrieval (MDR) databases. The estimates of school-related characteristics were obtained from the November 2019 Market Data Retrieval information. The MDR database contains the most recent data on schools, some of which may not be reflected in the NCES data. These data can be directly linked to assessment data of students included in the norming sample.

Table 19 on page 51 shows the percentages of students in grades 1–8 by region, school enrollment, school socioeconomic status, location, and school type nationally, and for the fall and spring norming samples. There were some missing data for some students where MDR data could not be linked to the student assessment data. For the fall norming sample 15.83% of the sample was missing MDR data, and for the spring norming sample 21.19% of the sample was missing MDR data.

A brief description of the geographic region, school enrollment, school socioeconomic, location and school type variables based on MDR is provided below.

Geographic Region

Using the categories established by the National Center for Education Statistics (NCES), students were grouped into four geographic regions as defined below: Northeast, Southeast, Midwest, and West.

Northeast

Connecticut, District of Columbia, Delaware, Massachusetts, Maryland, Maine, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont

Southeast

Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, West Virginia

Midwest

Iowa, Illinois, Indiana, Kansas, Minnesota, Missouri, North Dakota, Nebraska, Ohio, South Dakota, Michigan, Wisconsin

West

Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, New Mexico, Nevada, Oklahoma, Oregon, Texas, Utah, Washington, Wyoming

School Size

Based on total school enrollment, schools were classified into one of three school size groups: small schools had under 200 students enrolled, medium schools had 200–499 students enrolled, and large schools had 500 or more students enrolled.

Socioeconomic Status

Schools were classified into one of four classifications based on the percentage of students in the school who had free or reduced student lunch. The classifications were coded as follows:

- High socioeconomic status (0%–24%)
- Above-median socioeconomic status (25%–49%)
- Below-median socioeconomic status (50%–74%)
- Low socioeconomic status (75%–100%)

School Location

Schools were classified into one of four categories based on the school metro code type. The classifications were as follows:

- Rural
- Suburban
- Town
- Urban

School Type

Schools were also classified into one of two categories based on whether the school was a public or non-public school.

Table 19 presents the sample characteristic percentages for the MDR variables for the fall and spring norming samples.

Table 19: Sample Characteristics for Fall and Spring Norming Samples

		National Estimates	Fall Norming Samples	Spring Norming Samples
Region	Midwest	25.96%	9.42%	13.07%
	Northeast	19.41%	5.02%	4.78%
	Southeast	21.48%	3.70%	3.06%
	West	33.15%	81.86%	79.08%
School Enrollment	< 200	24.55%	0.44%	0.68%
	200–499	39.98%	21.99%	22.11%
	≥ 500	35.47%	77.57%	77.21%
District Socioeconomic Status	Low	12.16%	60.65%	63.41%
	Below Median	17.85%	23.77%	22.00%
	Above Median	19.35%	9.02%	7.91%
	High	50.64%	6.56%	6.69%
Location	Rural	23.64%	3.60%	3.10%
	Suburban	33.14%	38.95%	40.72%
	Town	14.35%	6.59%	5.87%
	Urban	28.87%	50.86%	50.30%
School Type	Public	82.08%	98.79%	98.49%
	Non-Public	17.92%	1.21%	1.51%

The norming sample also included students of different genders and ethnicities as well as students with disabilities and English Language Learners. Table 20 provides information on the demographic characteristics of students in the sample. No weighting was done based on these demographic variables; they are provided to help describe the sample of students and the schools they attended. Because Star assessment users can opt out of providing individual student demographic information such as gender and ethnicity/race, some students were missing demographic data; the sample summaries in Table 20 are based on only those students for whom gender and ethnicity information were available. Data on students with disabilities and English Language Learners are not provided because many Star assessment users do not enter that information, and initial analyses of data in the norming samples suggested that the percentages of students with disabilities and English Languages may underestimate the total percentage of students in these two groups. School type was defined to be either public (including charter schools) or non-public (private, Catholic).

The most recent data on student demographics is from NCES 2018–2019 data (http://nces.ed.gov/ccd/elsi/) and the NCES Private School Universe Survey (PSS) 2017–2018. For Non-Public, the PSS percent of 0.80% for Pacific Islander was combined with the 6.50% for Asian to equal 7.30% in Table 20 below.

NCES reports in 2019 approximately 56.6 million students attended elementary and secondary schools in the US; 50.8 million (90%) were in public schools and 5.8 million (10%) were in non-public schools. NCES does not report gender balance for public and non-public schools. In 2017, 9.6% of public school students were learning English as a second language.

Table 20: Student Demographics and School Information: Sample Percentages for Fall and Spring Norming Samples

			National Estimate	Fall Norming Sample	Spring Norming Sample
Gender	Public	Female	49.5%	49.99%	50.05%
		Male	50.5%	50.01%	49.95%
	Non-Public	Female	—	55.45%	56.64%
		Male	—	44.55%	43.36%
Race/Ethnicity	Public	American Indian	0.97%	3.82%	3.25%
		Asian	5.64%	0.59%	0.41%
		Black	15.15%	1.68%	1.56%
		Hispanic	27.12%	83.47%	85.16%
		White	47.08%	8.98%	8.02%
		Multiple Race[a]	4.05%	1.46%	1.6%
	Non-Public	American Indian	0.50%	—	0.33%
		Asian	7.30%	4.43%	3.93%
		Black	9.30%	1.3%	0.98%
		Hispanic	11.30%	43.23%	37.38%
		White	66.70%	42.71%	51.48%
		Multiple Race[a]	4.90%	8.33%	5.9%

a. Students identified as belonging to two or more races.

Test Administration

All students took the current version of the Star Reading Spanish tests under normal administration procedures. Some students in the norming sample took the assessment two or more times within the norming windows; scores from their initial test administration in the fall and the last test administration in the spring were used for computing the norms.

Data Analysis

Student test records were compiled from the complete database of Star Reading Spanish test users. Data were from the 2018–2019 school year from August to June. Students' unified scale Rasch scores on their first Star Reading Spanish test taken during the first and second months of the school year based on grade placement were used to compute norms for the fall; students' Rasch scores on the last Star Reading Spanish test taken during the 7th and 8th months of the school year were used to compute norms for the spring. Interpolation was used to estimate norms for times of the year between the first month in the fall and the last month in the spring. The norms were based on the distribution of Rasch scores for each grade.

Table 21 provides descriptive statistics for each grade with respect to the normative sample performance, in the Unified scaled score units.

Table 21: Descriptive Statistics for Scaled Scores by Grade for the Norming Sample on the Unified Scale

	Fall Unified Scaled Scores				Spring Unified Scaled Scores			
Grade	N	Mean	Standard Deviation	Median	N	Mean	Standard Deviation	Median
1	6,730	776	46	764	10,340	833	64	834
2	22,180	832	60	826	20,710	876	69	880
3	21,600	876	65	877	18,610	907	72	910
4	17,410	912	68	913	14,790	935	74	938
5	13,020	938	71	940	10,380	951	78	956
6	5,200	954	74	959	3,990	962	81	968
7	3,290	973	78	979	2,600	989	85	996
8	3,320	991	81	1000	2,110	1000	87	1008

Growth Norms

Student achievement typically is thought of in terms of status: a student's performance at one point in time. However, this ignores important information about a student's learning trajectory—how much students are growing over a period of time. When educators are able to consider growth information—the amount or rate of change over time—alongside current status, a richer picture of the student emerges, empowering educators to make better instructional decisions.

To facilitate deeper understanding of achievement, Renaissance Learning maintains growth norms for Star adaptive Assessments that provide insight both on growth to date and likely growth in the future. Growth norms are currently available for both English and Spanish versions of Star Math, Star Reading, and Star Early Literacy.

The growth model used by Star Assessments is Student Growth Percentile (Betebenner, 2009). SGPs were developed by Dr. Damian Betebenner, originally in partnership with several state departments of education.[1] It should be noted that the initial development of SGP involved annual state summative tests with reasonably constrained testing periods within each state. Because Star tests may be taken at multiple times throughout the year, a number of adaptations to the original model were made. For more information about Star SGPs, please refer to this overview: http://doc.renlearn.com/KMNet/R00571375CF86BBF.pdf

SGPs are norm-referenced estimates that compare a student's growth to that of his or her academic peers nationwide. Academic peers are defined as those students in the same grade[2] with a similar score history. SGPs are generated via a process that uses quantile regression to provide a measure of how much a student changed from one Star testing window to the next relative to other students with similar score histories.

SGPs range from 1–99 and are interpreted similarly to Percentile Ranks, with 50 indicating typical or expected growth. For instance, an SGP score of 37 means that a student grew as much or more than 37 percent of her academic peers, and less than about 63 percent of her academic peers. The Star SGP package also produces a range of future growth estimates. Those are mostly hidden from users but are presented in goal-setting and related applications to help users understand what typical or expected growth looks like for a given student. At present, the Star Reading Spanish SGP growth norms are based on a sample of about 375,000 student records across grades 1–8.

1. Core SGP documentation and source code are publicly available at https://cran.r-project.org/web/packages/SGP/index.html.
2. In rare instances, for some grade and testing window combinations, data may be pooled across nearby grades in order to increase sample sizes.

Score Definitions

This chapter enumerates the scores reported by Star Reading Spanish, including scaled scores, norm-referenced, and criterion-referenced scores.

Types of Test Scores

In a broad sense, Star Reading Spanish software provides three different types of test scores that measure student performance in different ways:

- *Scaled scores.* Star Reading Spanish creates a virtually unlimited number of test forms as it dynamically interacts with the students taking the test. In order to make the results of all tests comparable, and in order to provide a basis for deriving the other types of test scores described below, it is necessary to convert the results of Star Reading Spanish tests to scores on a common scale. Star Reading Spanish software does this in two steps. First, maximum likelihood is used to estimate each student's score on the Rasch ability scale, based on the difficulty of the items administered and the pattern of right and wrong answers. In the case that a student gets all items right or wrong, a proprietary Bayesian-modal item response theory estimation method is used. Second, the Rasch ability scores are converted to scaled scores. The score scale on which the scaled scores are reported is known as the "Unified" score scale.

  Unified Scale Scores

  Renaissance developed a single score scale that applies to all Star assessments: the Unified score scale. That development began with equating each test's underlying Rasch ability scales to a common Rasch scale; the result was the "unified Rasch scale," which is an extension of the Rasch scale used in Star Reading. The next step was to develop an integer scale based on the unified Rasch scale, with scale scores anchored to important points on the original Enterprise score scales that were developed for Star Math and Star Reading. The end result was a reported score scale that extends from 200 to 1400.

  Star Math, Star Reading, Star Reading Spanish, and Star Math Spanish Unified report scale scores that range from 600 to 1400. Star Early Literacy and Star Early Literacy Spanish Unified reported scale scores range from 200 to 1100. One benefit of the Unified scale is an improvement in certain properties of the scale scores: test scores are much less variable from grade to grade; measurement error is likewise

less variable; and Unified score reliability is slightly higher than that of the Enterprise scores. The Unified score scale is the only scale used to report results for Star Spanish assessments.

- *Criterion-referenced scores* describe a student's performance relative to a specific content domain or to a standard. Such scores may be expressed either on a continuous score scale or as a classification. An example of a criterion-referenced score on a continuous scale is a percent-correct score, which expresses what proportion of test questions the student can answer correctly in the content domain. An example of a criterion-referenced classification is a proficiency category on a standards-based assessment: the student may be said to be "proficient" or not, depending on whether the student's score equals, exceeds, or falls below a specific criterion (the "standard") used to define "proficiency" on the standards-based test. The criterion-referenced score reported by Star Reading Spanish is the Instructional Reading Level, which compares a student's test performance to the Rasch difficulty distributions of discrete sets of Star Reading Spanish test items identified with grade levels from 1 to 8. The Instructional Reading Level is the highest grade level at which the student is estimated to comprehend 80 percent of the text written at that level.
- *Norm-referenced scores* compare a student's test results to the results of other students who have taken the same test. In this case, scores provide a relative measure of student achievement compared to the performance of a reference group of students at a given time. Percentile Ranks and Grade Equivalents are the two primary norm-referenced scores available in Star Reading Spanish software. Both of these scores are based on a comparison of a student's test results to the data collected during the 2018 Star Reading Spanish norming program.

Grade Equivalent (GE)

A Grade Equivalent (GE) indicates the grade placement of students for whom a particular score is typical. If a student receives a GE of 8.0, this means that the student scored as well on Star Reading Spanish as did the typical student at the beginning of grade 8 in the norming sample. It does not necessarily mean that the student can read independently at an eighth-grade level—only that he or she obtained a Scaled Score as high as the average eighth-grade student in the norms group.

GE scores are often misinterpreted as though they convey information about what a student knows or can do—that is, as if they were criterion-referenced scores. To the contrary, GE scores are norm-referenced.

Star Reading Spanish Grade Equivalents range from 1.0–8.9. The scale divides the academic year into 10 monthly increments and is expressed as a decimal with the unit denoting the grade level and the individual "months" in tenths. Table 22 indicates how the GE scale corresponds to the various calendar months. For example, if a student obtained a GE of 4.6 on a Star Reading Spanish assessment, this would suggest that the student was performing similarly to the average student in the fourth grade at the sixth month (March) of the academic year. Because Star Reading Spanish norms are based on fall and spring score data only, monthly GE scores are derived through interpolation by fitting a curve to the grade-by-grade medians. Table 23 on page 69 contains the Scaled Score to GE conversions.

Table 22: Incremental Grade Placements per Month

Month	Decimal Increment	Month	Decimal Increment
July	0.00 or 0.99[a]	January	0.4
August	0.00 or 0.99[a]	February	0.5
September	0.0	March	0.6
October	0.1	April	0.7
November	0.2	May	0.8
December	0.3	June	0.9

a. Depends on the current school year set in Renaissance.

The Grade Equivalent scale is not an equal-interval scale. For example, an increase of 50 Scaled Score points might represent only two or three months of GE change at the lower grades, but over a year of GE change in higher grades. This is because student growth in reading (and other academic areas) is not linear; it occurs much more rapidly in the lower grades and slows greatly after the middle years. Consideration of this should be made when averaging GE scores, especially if it is done across two or more grades.

Comparing the Star Reading Spanish Test with Conventional Tests

Because the Star Reading Spanish test adapts to the reading level of the student being tested, Star Reading Spanish GE scores are more consistently accurate across the achievement spectrum than those provided by conventional test instruments. Grade Equivalent scores obtained using conventional (non-adaptive) test instruments are less accurate when a

student's grade placement and GE score differ markedly. It is not uncommon for a fourth-grade student to obtain a GE score of 8.9 when using a conventional test instrument. However, this does not necessarily mean that the student is performing at a level typical of an end-of-year eighth-grader. More likely, it means that the student answered all, or nearly all, of the items correctly on the conventional test and thus performed beyond the range of the fourth-grade test.

Star Reading Spanish Grade Equivalent scores are more consistently accurate—even as a student's achievement level deviates from the level of grade placement. A student may be tested on any level of material, depending upon his or her actual performance on the test; students are tested on items of an appropriate level of difficulty, based on their individual level of achievement. Thus, a GE score of 7.6 indicates that the student's score can be appropriately compared to that of a typical seventh-grader in the sixth month of the school year (with the same caveat as before—it does not mean that the student can actually handle seventh-grade reading material).

Instructional Reading Level (IRL)

The Instructional Reading Level is a criterion-referenced score that indicates the highest reading level at which the student can effectively be taught. In other words, IRLs tell you the reading level at which students can recognize words and comprehend written instructional material with some assistance. A sixth-grade student with an IRL of 4.0, for example, would be best served by instructional materials prepared at the fourth-grade level. IRLs are represented by either numbers or letters indicating a particular grade. For Star Reading Spanish, number codes represent IRLs for grades 1.0–8.9. IRL letter codes include PP (Pre-Primer) and P (Primer, grades 0.1–0.9).

As a construct, instructional reading levels have existed in the field of reading education for over seventy years. During this time, a variety of assessment instruments have been developed using different measurement criteria that teachers can use to estimate IRL. Star Reading Spanish software determines the Instructional Reading Level by comparing a student's test performance to the Rasch difficulty distributions of discrete sets of Star Reading Spanish test items identified with reading grade levels from 1 to 8. The Instructional Reading Level is defined as the highest reading level at which the student can read at 90–98 percent word recognition (Gickling & Haverape, 1981; Johnson, Kress & Pikulski, 1987; McCormick, 1999) and with 80 percent comprehension or higher (Gickling & Thompson, 2001). Although Star Reading Spanish does not directly assess word recognition, Star Reading Spanish uses the student's

Rasch ability scores, in conjunction with the Rasch difficulty parameters of graded Spanish reading items, to determine the proportion of items a student can comprehend at each grade level.

Special IRL Scores

If a student's performance on Star Reading Spanish indicates an IRL below the first grade, Star Reading Spanish software will automatically assign an IRL score of Primer (P) or Pre-Primer (PP). Because the kindergarten-level test items are designed so that even readers of very early levels can understand them, a Primer or Pre-Primer IRL means that the student is essentially a non-reader. There are, however, other unusual circumstances that could cause a student to receive an IRL of Primer or Pre-Primer. Most often, this happens when a student simply does not try or purposely answers questions incorrectly.

Understanding IRL and GE Scores

One strength of Star Reading Spanish software is that it provides both criterion-referenced and norm-referenced scores. As such, it provides more than one frame of reference for describing a student's current reading performance. The two frames of reference differ significantly, however, so it is important to understand the two estimates and their development when making interpretations of Star Reading Spanish results.

The Instructional Reading Level (IRL) is a criterion-referenced score. It provides an estimate of the grade level of written material with which the student can most effectively be taught. While the IRL, like any test result, is simply an estimate, it provides a useful indication of the level of material on which the student should be receiving instruction. For example, if a student (regardless of current grade placement) receives a Star Reading Spanish IRL of 4.0, this indicates that the student can most likely learn without experiencing too many difficulties when using materials written to be on a fourth-grade level.

The IRL is estimated based on the student's pattern of responses to the Star Reading Spanish items. A given student's IRL is the highest grade level of items at which it is estimated that the student can correctly answer at least 80 percent of the items.

In effect, the IRL references each student's Star Reading Spanish performance to the difficulty of written material appropriate for instruction. This is a

valuable piece of information in planning the instructional program for individuals or groups of students.

The Grade Equivalent (GE) is a norm-referenced score. It provides a comparison of a student's performance with that of other students around the nation. If a student receives a GE of 4.0, this means that the student scored as well on the Star Reading Spanish test as did the typical student at the beginning of grade 4. It does not mean that the student can read books that are written at a fourth-grade level—only that he or she reads as well as fourth-grade students in the norms group.

In general, IRLs and GEs will differ. These differences are caused by the fact that the two score metrics are designed to provide different information. That is, IRLs estimate the level of text that a student can read with some instructional assistance; GEs express a student's performance in terms of the grade level for which that performance is typical. Usually, a student's GE score will be higher than the IRL.

The score to be used depends on the information desired. If a teacher or educator wishes to know how a student's Star Reading Spanish score compares with that of other students in the norms group, either the GE or the Percentile Rank should be used. If the teacher or educator wants to know what level of instructional materials a student should be using for ongoing classroom schooling, the IRL is the preferred score. Again, both scores are estimates of a student's current level of reading achievement. They simply provide two ways of interpreting this performance—relative to a national sample of students (GE) or relative to the level of written material the student can read successfully (IRL).

Percentile Rank (PR)

Percentile Rank is a norm-referenced score that indicates the percentage of students in the same grade and at the same point of time in the school year who obtained scores lower than the score of a particular student. In other words, Percentile Ranks show how an individual student's performance compares to that of his or her same-grade peers in the national norming sample. For example, a Percentile Rank of 85 means that the student is performing at a level that exceeds 85 percent of other students in that grade at the same time of the year. Percentile Ranks simply indicate how a student performed compared to the others in the Star Reading Spanish norms group. The range of Percentile Ranks is 1–99.

The Percentile Rank scale is not an equal-interval scale. For example, for a student with a grade placement of 7.0, a Scaled Score of 1,059 corresponds to a PR of 80, and a Scaled Score of 1,091 corresponds to a PR of 90. Thus, a difference of 32 Scaled Score points represents a 10-point difference in PR. However, for students at the same 7.0 grade placement, a Scaled Score of 991 corresponds to a PR of 50, and a Scaled Score of 1,011 corresponds to a PR of 60. While there is now only a 10-point difference in Scaled Scores, there is still a 10-point difference in PR. For this reason, PR scores should not be averaged or otherwise algebraically manipulated. NCE scores are much more appropriate for these activities.

Table 24 on page 72 contains an abridged version of the Scaled Score to Percentile Rank conversion table that the Star Reading Spanish software uses. The actual table includes data for all of the monthly grade placement values from 1.0–8.9. Because the norming of Star Reading Spanish occurred in the fall and the spring, the first-month and last-month are empirically based, and the remaining monthly values were estimated by interpolating between the empirical points for the Fall and Spring norms.

Normal Curve Equivalent (NCE)

Normal Curve Equivalents (NCEs) are scores that have been scaled in such a way that they have a normal distribution, with a mean of 50 and a standard deviation of 21.06 in the normative sample for a given test. Because they range from 1–99, they appear similar to Percentile Ranks, but they have the advantage of being based on an equal interval scale. That is, the difference between two successive scores on the scale has the same meaning throughout the scale. NCEs are useful for purposes of statistically manipulating norm-referenced test results, such as interpolating test scores, calculating averages, and computing correlation coefficients between different tests. For example, in Star Reading Spanish score reports, average Percentile Ranks are obtained by first converting the PR values to NCE values, averaging the NCE values, and then converting the average NCE back to a PR.

Table 25 on page 75 provides the NCEs corresponding to integer PR values and facilitates the conversion of PRs to NCEs. Table 26 on page 76 provides the conversions from NCE to PR. The NCE values are given as a range of scores that convert to the corresponding PR value.

Student Growth Percentile (SGP)

Student Growth Percentiles (SGPs) are a norm-referenced quantification of individual student growth derived using quantile regression techniques. An SGP compares a student's growth to that of his or her academic peers nationwide with a similar achievement history on Star assessments. Academic peers are students who

- are in the same grade,[1]
- had the same scores on the current test and (up to) two prior tests from different testing windows, and
- took the most recent test and the first prior test on the same dates.

SGPs provide a measure of how a student changed from one Star testing window[2] to the next relative to other students with similar starting Star Reading Spanish scores. SGPs range from 1–99 and interpretation is similar to that of Percentile Rank scores; lower numbers indicate lower relative growth and higher numbers show higher relative growth. For example, an SGP of 70 means that the student's growth from one test window to another exceeds the growth of 70% of students nationwide in the same grade with a similar Star Reading Spanish score history. All students, no matter their starting Star score, have an equal chance to demonstrate growth at any of the 99 percentiles.

SGPs are often used to indicate whether a student's growth is more or less than can be expected. For example, without an SGP, a teacher would not know if a Scaled Score increase of 100 points represents good, not-so-good, or average growth. This is because students of differing achievement levels in different grades grow at different rates relative to the Star Reading Spanish scale. For example, a high-achieving second-grader grows at a different rate than a low-achieving second-grader. Similarly, a high-achieving second-grader grows at a different rate than a high-achieving eighth-grader.

SGPs can be aggregated to describe typical growth for groups of students—for example, a class, grade, or school as a whole—by calculating the group's median, or middle, growth percentile. No matter how SGPs are aggregated, whether at the class, grade, or school level, the statistic and its interpretation remain the same. For example, if the students in one class have a median SGP of 62, that particular group of students, on average, achieved higher growth than their academic peers.

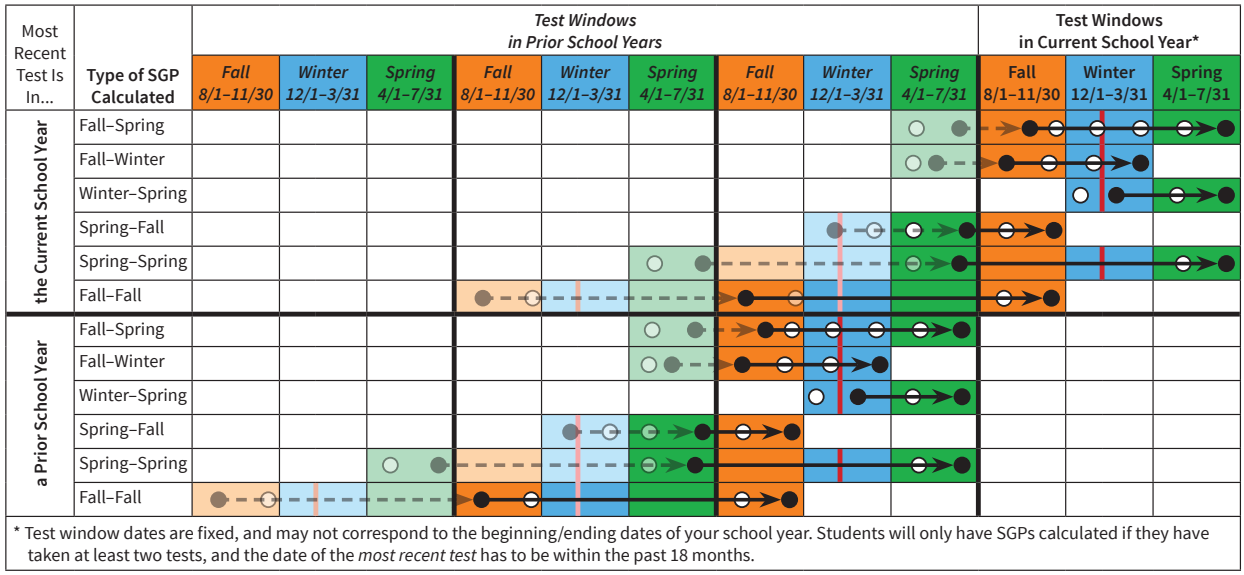
---

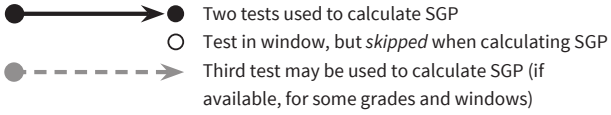
1 In rare instances, for some grade and testing window combinations, data may be pooled across nearby grades in order to increase sample sizes.
2. We collect data for our growth norms during three different time periods: fall, winter, and spring. More information about these time periods is provided on page 66.

SGP is calculated for students who have taken at least two tests (a *current* test and a *prior* test) within at least two different testing windows (Fall, Winter, or Spring).

If a student has taken more than one test in a single test window, the SGP calculation is based off the following tests:

- The current test is always the last test taken in a testing window.
- The test used as the prior test depends on what testing window it falls in:
  - Fall window: The first test taken in the Fall window is used.
  - Winter window: The test taken closest to January 15 in the Winter window is used.
  - Spring window: The last test taken in the Spring window is used.





Test Window	If more than one test was taken in a prior test window, which is used to calculate SGP?
Fall Window	First test taken
Winter Window	Test closest to 1/15 (red line)
Spring Window	Last test taken

Grade Placement

Star Reading Spanish software uses the student's grade placement—grade and month of the school year—when determining the norm-referenced scores. The values of PR and NCE are based not only on what scaled score the student achieved but also on the grade placement of the student at the time of the test (for example, a second-grader in the seventh month with a scaled score of 889 would have a PR of 65, while a third-grader in the seventh month with the same scaled score would have a PR of 46). Thus, it is crucial that student records indicate the proper grade placement when students take a Star Reading Spanish test, and that any testing in July or August reflects the proper understanding of how Star Reading Spanish software deals with these months in determining grade placement.

Indicating the Appropriate Grade Placement

The numeric representation of a student's grade placement is based on the specific month and day in which he or she takes a test. Although teachers indicate a student's grade level using whole numbers, Star Reading Spanish software automatically adds fractional increments to that grade level based on the month and day of the test. To determine the appropriate increment, Star Reading Spanish software considers the standard school year to run from September—June and assigns increment values of 0.0–0.9 to these months.

Table 22 on page 60 summarizes the increment values assigned to each month.

The increment values for July and August depend on the school year setting:

- If teachers will use the July and August test scores to evaluate the student's reading performance at the beginning of the year, in the Renaissance program, make sure the start date for that school year is before your testing in July and August. Grades are automatically increased by one level in each successive school year, so promoting students to the next grade is not necessary. In this case, the increment value for July and August is 0.00 because these months are at the beginning of the school year.
- If teachers will use the test scores to evaluate the student's reading performance at the end of the school year, make sure the end date for that school year falls after your testing in July and August. In this case, the increment value for July and August is 0.99 because these months are at the end of the school year that has passed.

In addition to the tenths digit appended to the grade level to denote the month of the standard school year in which a test was taken, Star Reading Spanish appends a hundredths digit to denote the day on which a test was taken as well. The hundredths digit represents the fractional portion of a 30-day month. For example, the increment for a test taken on the sixth day of the month is 0.02. For a test taken on the twenty-fourth day of the month, the increment is 0.08.

If a school follows the standard school calendar used in Star Reading Spanish software and does not test in the summer, assigning the appropriate grade placements for students is relatively easy. However, if students will be tested in July or August—whether it is for a summer reading program or because the normal calendar extends into these months—grade placements become an extremely important issue.

To ensure the accurate determination of norm-referenced scores when testing in the summer, it must be determined when to set the next school year as the current school year, and thereby advance students from one grade to the next. In most cases, the guidelines above can be used.

Instructions for specifying school years and grade assignments can be found at https://help.renaissance.com/RP and https://help2.renaissance.com/setup.

Compensating for Incorrect Grade Placements

Teachers cannot make retroactive corrections to a student's grade placement by editing the grade assignments in a student's record or by adjusting the increments for the summer months after students have tested. In other words, Star Reading Spanish software cannot go back in time and correct scores resulting from erroneous grade placement information. Thus, it is extremely important for the test administrator to make sure that the proper grade placement procedures are being followed.

Conversion Tables

Table 23: Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score	
	Low	High
1	600	767
1.1	768	774
1.2	775	782
1.3	783	789
1.4	790	795
1.5	796	802
1.6	803	808
1.7	809	815
1.8	816	821
1.9	822	826
2	827	832
2.1	833	838
2.2	839	843
2.3	844	848
2.4	849	853
2.5	854	858
2.6	859	863
2.7	864	867
2.8	868	872
2.9	873	876
3	877	880
3.1	881	884
3.2	885	888
3.3	889	891
3.4	892	895
3.5	896	898
3.6	899	902
3.7	903	905

Table 23: Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score	
	Low	High
3.8	906	908
3.9	909	911
4	912	914
4.1	915	917
4.2	918	919
4.3	920	922
4.4	923	924
4.5	925	927
4.6	928	929
4.7	930	932
4.8	933	934
4.9	935	936
5	937	938
5.1	939	941
5.2	942	943
5.3	944	945
5.4	946	947
5.5	948	949
5.6	950	951
5.7	952	953
5.8	954	955
5.9	956	957
6	958	959
6.1	960	961
6.2	962	963
6.3	964	965
6.4	966	967
6.5	968	968
6.6	969	970
6.7	971	972
6.8	973	974

Table 23: Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score	
	Low	High
6.9	975	976
7	977	978
7.1	979	980
7.2	981	982
7.3	983	984
7.4	985	986
7.5	987	988
7.6	989	990
7.7	991	992
7.8	993	994
7.9	995	995
8	996	997
8.1	998	999
8.2	1000	1001
8.3	1002	1002
8.4	1003	1004
8.5	1005	1005
8.6	1006	1007
8.7	1008	1008
8.8	1009	1009
8.9	1010	1010
> 8.9	1011	1400

Table 24: Scaled Score to Percentile Ranks Conversion by Grade on the Unified Scale

	Grade (First Month)							
PR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	600	600	600	600	600	600	600	600
2	709	736	757	779	793	806	805	817
3	715	742	764	786	802	814	814	828
4	718	746	767	792	809	821	824	838
5	720	750	772	797	815	826	831	844
6	724	752	775	801	822	832	840	849
7	726	754	778	806	826	838	845	856
8	727	757	781	811	832	844	852	862
9	728	759	784	814	837	849	858	868
10	731	760	787	819	841	853	864	873
11	732	762	790	823	846	857	870	878
12	733	764	792	827	850	861	875	884
13	734	766	795	830	854	866	879	890
14	735	767	798	834	857	869	883	896
15	736	768	800	838	861	872	886	900
16	738	770	803	841	865	875	890	904
17	739	772	806	845	867	878	894	909
18	-	773	809	848	871	881	898	912
19	740	774	812	851	874	885	902	916
20	741	776	815	854	876	888	906	922
21	742	778	817	856	879	891	909	925
22	743	779	820	859	881	894	913	929
23	744	780	823	861	883	896	917	932
24	745	781	826	864	886	899	919	935
25	-	783	828	866	888	902	921	939
26	746	785	831	868	890	905	923	942
27	-	786	833	870	892	908	925	946
28	747	787	835	872	895	910	928	949
29	748	789	838	874	897	913	931	952
30	749	790	840	875	899	916	933	956
31	750	792	842	878	901	918	936	958
32	751	793	845	880	904	920	939	961
33	-	795	847	882	906	922	942	964
34	752	797	849	883	908	925	944	966

Table 24: Scaled Score to Percentile Ranks Conversion by Grade on the Unified Scale

	Grade (First Month)							
PR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
35	-	798	851	885	910	927	947	969
36	753	800	853	887	912	929	950	971
37	-	801	855	889	915	931	953	973
38	754	803	857	891	916	933	956	975
39	755	805	859	893	918	935	958	978
40	756	807	860	895	920	937	960	980
41	757	809	862	897	922	939	962	982
42	758	811	864	898	924	943	965	984
43	-	812	866	900	926	945	967	986
44	759	814	868	902	928	947	969	988
45	-	816	869	904	930	949	970	990
46	760	818	871	906	932	951	972	992
47	761	820	872	908	934	953	974	994
48	762	822	874	910	936	955	976	997
49	763	824	876	911	938	957	978	998
50	764	826	877	913	940	959	979	1000
51	765	828	879	915	942	961	981	1002
52	-	830	880	917	944	963	983	1003
53	766	832	882	918	946	965	985	1005
54	767	834	884	920	948	966	987	1007
55	768	836	885	922	950	968	989	1010
56	769	838	887	924	952	970	990	1011
57	-	840	889	925	954	972	992	1013
58	771	842	891	927	956	973	994	1015
59	772	844	892	929	958	975	996	1016
60	773	846	894	931	960	977	999	1018
61	774	848	896	933	962	979	1000	1020
62	775	850	898	935	964	981	1002	1022
63	776	852	900	937	966	982	1005	1023
64	777	854	902	939	968	985	1007	1025
65	778	856	904	941	970	987	1009	1027
66	779	858	906	943	972	989	1011	1029
67	781	859	907	945	974	992	1013	1031
68	782	861	909	947	976	994	1015	1033

Table 24: Scaled Score to Percentile Ranks Conversion by Grade on the Unified Scale

	Grade (First Month)							
PR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
69	784	863	911	949	978	996	1016	1035
70	785	865	912	951	979	999	1018	1037
71	786	867	914	954	981	1000	1021	1039
72	788	869	916	955	983	1003	1023	1042
73	790	870	918	957	985	1004	1024	1044
74	792	872	920	959	987	1006	1027	1046
75	794	874	922	961	990	1008	1029	1048
76	796	876	924	963	992	1010	1032	1050
77	798	878	926	965	994	1012	1034	1053
78	801	880	928	967	997	1014	1036	1056
79	803	882	930	970	999	1016	1039	1059
80	806	885	933	972	1001	1018	1040	1061
81	809	887	936	974	1004	1020	1042	1062
82	812	889	939	977	1007	1022	1045	1065
83	815	892	941	980	1009	1024	1048	1067
84	819	894	943	982	1012	1027	1051	1070
85	823	897	946	985	1014	1030	1054	1073
86	827	900	949	988	1017	1033	1058	1076
87	831	903	952	990	1019	1037	1060	1078
88	835	906	954	994	1022	1040	1064	1082
89	840	909	958	997	1025	1043	1067	1085
90	845	913	961	1001	1028	1047	1070	1089
91	849	916	965	1006	1031	1052	1073	1092
92	854	920	969	1009	1036	1057	1077	1096
93	860	925	973	1014	1040	1061	1081	1101
94	865	931	978	1019	1045	1066	1086	1108
95	872	937	982	1023	1051	1072	1090	1113
96	879	943	989	1030	1059	1077	1097	1120
97	887	950	997	1036	1067	1082	1103	1130
98	899	961	1008	1045	1077	1090	1113	1139
99	914	978	1023	1062	1091	1108	1126	1163

Table 25: Percentile Rank to Normal Curve Equivalent Conversions

PR	NCE	PR	NCE	PR	NCE	PR	NCE
1	1.0	26	36.5	51	50.5	76	64.9
2	6.7	27	37.1	52	51.1	77	65.6
3	10.4	28	37.7	53	51.6	78	66.3
4	13.1	29	38.3	54	52.1	79	67.0
5	15.4	30	39.0	55	52.6	80	67.7
6	17.3	31	39.6	56	53.2	81	68.5
7	18.9	32	40.1	57	53.7	82	69.3
8	20.4	33	40.7	58	54.2	83	70.1
9	21.8	34	41.3	59	54.8	84	70.9
10	23.0	35	41.9	60	55.3	85	71.8
11	24.2	36	42.5	61	55.9	86	72.8
12	25.3	37	43.0	62	56.4	87	73.7
13	26.3	38	43.6	63	57.0	88	74.7
14	27.2	39	44.1	64	57.5	89	75.8
15	28.2	40	44.7	65	58.1	90	77.0
16	29.1	41	45.2	66	58.7	91	78.2
17	29.9	42	45.8	67	59.3	92	79.6
18	30.7	43	46.3	68	59.9	93	81.1
19	31.5	44	46.8	69	60.4	94	82.7
20	32.3	45	47.4	70	61.0	95	84.6
21	33.0	46	47.9	71	61.7	96	86.9
22	33.7	47	48.4	72	62.3	97	89.6
23	34.4	48	48.9	73	62.9	98	93.3
24	35.1	49	49.5	74	63.5	99	99.0
25	35.8	50	50.0	75	64.2		

Table 26: Normal Curve Equivalent to Percentile Rank Conversion

NCE Range			NCE Range			NCE Range			NCE Range		
Low	**High**	**PR**	**Low**	**High**	**PR**	**Low**	**High**	**PR**	**Low**	**High**	**PR**
1.0	4.0	1	36.1	36.7	26	50.3	50.7	51	64.6	65.1	76
4.1	8.5	2	36.8	37.3	27	50.8	51.2	52	65.2	65.8	77
8.6	11.7	3	37.4	38.0	28	51.3	51.8	53	65.9	66.5	78
11.8	14.1	4	38.1	38.6	29	51.9	52.3	54	66.6	67.3	79
14.2	16.2	5	38.7	39.2	30	52.4	52.8	55	67.4	68.0	80
16.3	18.0	6	39.3	39.8	31	52.9	53.4	56	68.1	68.6	81
18.1	19.6	7	39.9	40.4	32	53.5	53.9	57	68.7	69.6	82
19.7	21.0	8	40.5	40.9	33	54.0	54.4	58	69.7	70.4	83
21.1	22.3	9	41.0	41.5	34	54.5	55.0	59	70.5	71.3	84
22.4	23.5	10	41.6	42.1	35	55.1	55.5	60	71.4	72.2	85
23.6	24.6	11	42.2	42.7	36	55.6	56.1	61	72.3	73.1	86
24.7	25.7	12	42.8	43.2	37	56.2	56.6	62	73.2	74.1	87
25.8	26.7	13	43.3	43.8	38	56.7	57.2	63	74.2	75.2	88
26.8	27.6	14	43.9	44.3	39	57.3	57.8	64	75.3	76.3	89
27.7	28.5	15	44.4	44.9	40	57.9	58.3	65	76.4	77.5	90
28.6	29.4	16	45.0	45.4	41	58.4	58.9	66	77.6	78.8	91
29.5	30.2	17	45.5	45.9	42	59.0	59.5	67	78.9	80.2	92
30.3	31.0	18	46.0	46.5	43	59.6	60.1	68	80.3	81.7	93
31.1	31.8	19	46.6	47.0	44	60.2	60.7	69	81.8	83.5	94
31.9	32.6	20	47.1	47.5	45	60.8	61.3	70	83.6	85.5	95
32.7	33.3	21	47.6	48.1	46	61.4	61.9	71	85.6	88.0	96
33.4	34.0	22	48.2	48.6	47	62.0	62.5	72	88.1	91.0	97
34.1	34.7	23	48.7	49.1	48	62.6	63.1	73	91.1	95.4	98
34.8	35.4	24	49.2	49.7	49	63.2	63.8	74	95.5	99.0	99
35.5	36.0	25	49.8	50.2	50	63.9	64.5	75			

Note: Table 27 can be used as a reference, but keep in mind that there could be a slightly different IRL displayed on reports (+/– 0.1) depending on the student's Rasch score used to calculate a Scaled Score and IRL.

Table 27: Scaled Score to Instructional Reading Level Conversions

IRL		Unified Scaled Score	
		Low	High
Pre-Primer (PP): < 0		600	825
Primer (P): 0.0–0.9	0.0	826	828
	0.1	829	832
	0.2	833	835
	0.3	836	839
	0.4	840	842
	0.5	843	846
	0.6	847	849
	0.7	850	854
	0.8	855	856
	0.9	857	861
1		862	863
1.1		864	868
1.2		869	871
1.3		872	873
1.4		874	878
1.5		879	880
1.6		881	885
1.7		886	887
1.8		888	892
1.9		893	895
2		896	899
2.1		900	902
2.2		903	906
2.3		907	909
2.4		910	913
2.5		914	916
2.6		917	920
2.7		917	920

Table 27: Scaled Score to Instructional Reading Level Conversions

IRL	Unified Scaled Score	
	Low	High
2.8	921	923
2.9	924	927
3	928	930
3.1	931	933
3.2	934	937
3.3	938	940
3.4	941	944
3.5	945	947
3.6	948	950
3.7	951	954
3.8	955	957
3.9	958	961
4	962	964
4.1	965	968
4.2	969	971
4.3	972	974
4.4	975	978
4.5	979	981
4.6	982	985
4.7	986	988
4.8	989	992
4.9	993	995
5	996	999
5.1	1000	1002
5.2	1003	1006
5.3	1007	1009
5.4	1010	1014
5.5	1015	1016
5.6	1017	1021
5.7	1022	1023
5.8	1024	1028
5.9	1029	1030

Table 27: Scaled Score to Instructional Reading Level Conversions

IRL	Unified Scaled Score	
	Low	High
6	1031	1035
6.1	1036	1038
6.2	1039	1042
6.3	1043	1045
6.4	1046	1047
6.5	1048	1052
6.6	1053	1055
6.7	1056	1059
6.8	1060	1062
6.9	1063	1066
7	1067	1069
7.1	1070	1073
7.2	1074	1076
7.3	1077	1080
7.4	1081	1083
7.5	1084	1087
7.6	1088	1090
7.7	1091	1093
7.8	1094	1097
7.9	1098	1100
8	1101	1104
8.1	1105	1107
8.2	1108	1111
8.3	1112	1114
8.4	1115	1117
8.5	1118	1121
8.6	1122	1124
8.7	1125	1128
8.8	1129	1131
8.9	1132	1134
> 8.9	1135	1400

References

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). easyCBM® online progress monitoring assessment system. http://easycbm.com. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice,* 28(4), 42–51.

Gickling, E. E., & Havertape, S. (1981). *Curriculum-based assessment (CBA).* Minneapolis, MN: School Psychology Inservice Training Network.

Gickling, E. E., & Thompson, V. E. (2001). Putting the learning needs of children first. In B. Sornson (Ed.). *Preventing early learning failure.* Alexandria, VA: ASCD.

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

Johnson, M. S., Kress, R. A., & Pikulski, J. J. (1987). *Informal reading inventories.* Newark, DE: International Reading Association.

McCormick, S. (1999). *Instructing students who have literacy problems* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Owen, R. J. (1969) A Bayesian approach to tailored testing. *Research Bulletin* 69–92. Princeton, N. J.: Educational Testing Service.

Owen, R. J. (1975) A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association,* 70, 351–356.

R Core Team (2018). R: *A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software,* 48(2), 1–36.

Index

F

G

H

I

K

L

M

N

P

R

S

T

U

V

W

About Renaissance

Renaissance® transforms data about how students learn into instruments of empowerment for classroom teachers, enabling them to guide all students to achieve their full potentials. Through smart, data-driven educational technology solutions that amplify teachers' effectiveness, Renaissance helps teachers teach better, students learn better, and school administrators lead better. By supporting teachers in the classroom but not supplanting them, Renaissance solutions deliver tight learning feedback loops: between teachers and students, between assessment of skills mastery and the resources to spur progress, and between each student's current skills and future academic growth.

RENAISSANCE®

 | (800) 338-4204 | www.renaissance.com

Fremont | London | Madison | Minneapolis | New York | Sydney | Toronto | Vancouver | Wisconsin Rapids

200608