Copyright Notice

Copyright © 2014 by Renaissance Learning, Inc. All Rights Reserved.

This publication is protected by US and international copyright laws. It is unlawful to duplicate or reproduce any copyrighted material without authorization from the copyright holder. This document may be reproduced only by staff members in schools that have a license for STAR Reading software. For more information, contact Renaissance Learning, Inc., at the address above.

All logos, designs, and brand names for Renaissance Learning’s products and services, including but not limited to Accelerated Math, Accelerated Reader, AccelScan, AccelTest, AR, ATOS, Core Progress, English in a Flash, Learnalytics, Progress Pulse, Renaissance Home Connect, Renaissance Learning, Renaissance Place, Renaissance-U, STAR, STAR 360, STAR Custom, STAR Early Literacy, STAR Math, STAR Reading, STAR Reading Spanish, Successful Reader, and Subtext, are trademarks of Renaissance Learning, Inc., and its subsidiaries, registered, common law, or pending registration in the United States and other countries. All other product and company names should be considered the property of their respective companies and organizations.

iPad and Macintosh are trademarks of Apple Inc., registered in the U.S. and other countries.
## Contents

### Introduction ...................................................... 1
- STAR Reading Spanish Purpose ........................................ 1
- Design of STAR Reading Spanish ........................................ 1
  - Overarching Design Considerations .................................. 1
- Test Interface .......................................................... 2
- Practice Session .......................................................... 2
- Adaptive Branching/Test Length ........................................ 3
  - Test Length ..................................................................... 3
- Test Repetition .............................................................. 4
- Item Time Limits ............................................................ 4
- Test Security ...................................................................... 5
  - Split-Application Model ................................................... 5
  - Individualized Tests ........................................................ 5
  - Data Encryption .................................................................. 6
  - Access Levels and Capabilities .......................................... 6
  - Test Monitoring/Password Entry ......................................... 6
  - Final Caveat ...................................................................... 7
- Test Administration Procedures ........................................... 7

### Content and Item Development ................................. 8
- Content Specification ...................................................... 8
  - The Spanish Vocabulary List ............................................. 8
- Item Development Specifications ......................................... 9
  - Vocabulary-in-Context Item Specifications ............................. 9
  - Level of Difficulty: Readability .......................................... 10
  - Level of Difficulty: Cognitive Load, Content Differentiation, and Presentation ........................................... 11
  - Efficiency in Use of Student Time ....................................... 11
  - Balanced Items: Bias and Fairness ....................................... 11
  - Accuracy of Content ........................................................ 12
  - Language Conventions ..................................................... 12

### Item and Scale Calibration ................................. 13
- Background ................................................................. 13
- Calibration of STAR Reading Spanish Items .......................... 13
- Item Difficulty ............................................................... 18
- Item Discrimination ......................................................... 18
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Response Function</td>
<td>19</td>
</tr>
<tr>
<td>Rules for Item Retention</td>
<td>20</td>
</tr>
<tr>
<td>Computer-Adaptive Test Design</td>
<td>21</td>
</tr>
<tr>
<td>Scoring in the STAR Reading Spanish Tests</td>
<td>22</td>
</tr>
<tr>
<td>Scale Calibration</td>
<td>22</td>
</tr>
<tr>
<td>Reliability and Measurement Precision</td>
<td>23</td>
</tr>
<tr>
<td>Generic Reliability</td>
<td>24</td>
</tr>
<tr>
<td>Split-Half Reliability</td>
<td>25</td>
</tr>
<tr>
<td>Standard Error of Measurement</td>
<td>26</td>
</tr>
<tr>
<td>Validity</td>
<td>28</td>
</tr>
<tr>
<td>STAR Reading Spanish Research Study Data</td>
<td>29</td>
</tr>
<tr>
<td>Relationship of STAR Reading Spanish to Skills Ratings</td>
<td>30</td>
</tr>
<tr>
<td>The Rating Instrument</td>
<td>30</td>
</tr>
<tr>
<td>Psychometric Properties of the Skills Ratings</td>
<td>32</td>
</tr>
<tr>
<td>Relationship of Scaled Scores to Skills Ratings</td>
<td>32</td>
</tr>
<tr>
<td>Relationship of STAR Reading Spanish Scores to Scores on Other Tests</td>
<td>35</td>
</tr>
<tr>
<td>of Spanish Reading Achievement</td>
<td>36</td>
</tr>
<tr>
<td>Summary of STAR Reading Spanish Validity Data</td>
<td>36</td>
</tr>
<tr>
<td>Score Definitions</td>
<td>37</td>
</tr>
<tr>
<td>Types of Test Scores</td>
<td>37</td>
</tr>
<tr>
<td>Criterion-Referenced Scores</td>
<td>37</td>
</tr>
<tr>
<td>Scaled Scores (SS)</td>
<td>37</td>
</tr>
<tr>
<td>Spanish Instruction Reading Level (IRL)</td>
<td>38</td>
</tr>
<tr>
<td>Special STAR Reading Spanish Scores</td>
<td>38</td>
</tr>
<tr>
<td>Spanish Zone of Proximal Development (ZPD)</td>
<td>38</td>
</tr>
<tr>
<td>STAR Reading Spanish in the Classroom</td>
<td>39</td>
</tr>
<tr>
<td>Formative Assessment</td>
<td>39</td>
</tr>
<tr>
<td>Measuring Growth</td>
<td>40</td>
</tr>
<tr>
<td>Absolute versus Relative Growth</td>
<td>40</td>
</tr>
<tr>
<td>Longitudinal Designs</td>
<td>40</td>
</tr>
<tr>
<td>Growth Estimates</td>
<td>41</td>
</tr>
<tr>
<td>References</td>
<td>43</td>
</tr>
<tr>
<td>Index</td>
<td>44</td>
</tr>
</tbody>
</table>
Introduction

STAR Reading Spanish Purpose

STAR Reading Spanish serves three purposes for students. First, it provides educators with quick and accurate estimates of reading comprehension using students’ Instructional Reading Levels. Second, it provides teachers with a ZPD (Zone of Proximal Development) to match students to books. Third, it provides the means for tracking growth in a consistent manner for all 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, brief test that builds confidence in their reading ability.
- For teachers, the STAR Reading Spanish test facilitates individualized instruction by identifying children who need remediation or enrichment most.
- For principals, STAR Reading Spanish provides regular, accurate reports on performance at the student, class, and grade level.
- For district administrators and assessment specialists, STAR Reading Spanish provides a reliable way to track reading growth at each school.

This manual documents the suitability of STAR Reading Spanish computer-adaptive testing for these purposes and demonstrates quantitatively how well this innovative instrument in reading assessment performs.

Design of STAR Reading Spanish

Overarching Design Considerations

The fundamentals of the STAR Reading Spanish design—computer-adaptive testing and vocabulary-in-context—were based on the years of experiences with STAR Reading and reviewed by Spanish experts to be valid for Spanish as well as English.

STAR Reading Spanish uses computer-adaptive testing. The primary advantage of using computer software to administer is the ability to tailor each student’s test based on his or her responses to previous items, making 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 paper-and-pencil tests because the test difficulty matches each individual’s performance level; students do not need 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 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 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.

The vocabulary-in-context test items, while using a common format for assessing reading, require reading comprehension. Each test item is a complete, contextual sentence with a tightly controlled vocabulary level. The semantics and syntax of each context sentence are arranged to provide clues as to the correct cloze word. The student must actually interpret the meaning of (in other words, comprehend) the sentence in order to choose the correct answer because all of the answer choices ‘fit’ the context sentence either semantically or syntactically. In effect, each sentence provides a mini-selection of which the student demonstrates the ability to interpret the correct meaning. This is, after all, what most reading theorists believe reading comprehension to be—the ability to draw meaning from text.

**Test Interface**

The STAR Reading Spanish test interface was designed to be both simple and effective. Students can use either the mouse or the keyboard to answer questions.

- If using the keyboard, students press one of the four 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 (Next) to enter the answer.

In April of 2013, the STAR Apps on iPad® was released, allowing students to take a STAR Reading Spanish test on an iPad®. Students tap the answer of choice and then tap Siguiente (Next) to enter the answer.

**Practice Session**

The practice session before the test 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 (out of up to seven
questions), 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 session with the student, the program will recommend that the teacher assist the student during the practice.

Adaptive Branching/Test Length

STAR Reading Spanish’s 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 the criterion-referenced score 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 86 percent of students will be able to answer the first item correctly. Teachers can override this typical value by entering an even lower Estimated Spanish Instructional Reading Level (Est. SP IRL) for the student. On the second and subsequent administrations, the STAR Reading Spanish 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 25 items of varying difficulty based on the student’s responses; this is sufficient information to obtain a reliable Scaled Score and determine the student’s Spanish Instructional Reading Level (SP IRL). 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 participating in the spring 2012 Research study. The results of the analysis of test completion time indicates that about half of the students at every grade will complete the STAR Reading Spanish test in less than 12 minutes, and even in the slowest grade (grade 1), 95 percent of students finished their STAR Reading Spanish test in less than 19 minutes.
Repeated testing allows teachers to measure a student’s reading growth over time. Renaissance Learning recommends administering STAR assessments two to three times a year for purposes of placement and growth. New students, or students for whom you occasionally need additional information, may be tested at any time. 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.

Table 1: Percentiles of Total Time to Complete the STAR Reading Spanish Assessment in the 2012 Research Study

<table>
<thead>
<tr>
<th>Grade</th>
<th>5th Percentile</th>
<th>25th Percentile</th>
<th>50th Percentile</th>
<th>75th Percentile</th>
<th>95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.23</td>
<td>8.27</td>
<td>11.05</td>
<td>13.95</td>
<td>18.78</td>
</tr>
<tr>
<td>2</td>
<td>4.52</td>
<td>7.33</td>
<td>9.42</td>
<td>12.15</td>
<td>16.48</td>
</tr>
<tr>
<td>3</td>
<td>5.23</td>
<td>7.33</td>
<td>9.22</td>
<td>11.25</td>
<td>15.40</td>
</tr>
<tr>
<td>4</td>
<td>4.92</td>
<td>6.77</td>
<td>8.32</td>
<td>10.77</td>
<td>16.08</td>
</tr>
<tr>
<td>5</td>
<td>4.15</td>
<td>6.23</td>
<td>7.45</td>
<td>8.98</td>
<td>11.80</td>
</tr>
</tbody>
</table>

Item Time Limits

The STAR Reading Spanish test has time-out limits for individual items that are based on a student’s grade level. Students in grades K–2 have up to 75 seconds to answer each item during their test sessions. Students in grades 3–12 are allowed 60 seconds to answer each vocabulary-in-context item. These time-out values are based on latency data obtained during item validation. No vocabulary-in-context items at any grade had latencies longer than 40 seconds on average. Thus, the time-out limit was set to 60 seconds for most students and increased to 75 seconds for the very young students.

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 students with disabilities. Test users who elect the extended time limit for their students should be aware that STAR Reading Spanish 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 three times longer than the standard time limits to answer each question. Therefore, students in grades K–2 with the extended time limit accommodation have up to 225 seconds to answer each item. Students in grades 3–12 with the extended time limit accommodation have 180 seconds to answer each vocabulary-in-context item.

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 clock 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 is 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

In the STAR Reading Spanish software, when students log in, they do not have access to the same functions that teachers, administrators, and other personnel can access. Students are allowed to test, but they have no other tasks available in STAR Reading Spanish; 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 Place program depends on the primary position assigned to that user and the capabilities the user has been granted in the Renaissance Place program. Each primary position is part of a user group. There are seven user groups: district administrator, district staff, school administrator, school staff, teacher, parent, and student. By default, each user group is granted a specific set of capabilities. Each capability corresponds to one or more tasks that can be performed in the program. The capabilities in these sets can be changed; capabilities can also be granted or removed on an individual level. Since users can be assigned to the district and/or one or more schools (and be assigned different primary positions at the different locations), and since the capabilities granted to a user can be customized, there are many levels of access an individual user can have.

Renaissance Place RT 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 on access and security, see the Renaissance Place Software Manual.

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 gives them access to the data and features that they are allowed based on their primary position and the capabilities that they have been granted. Personnel who log in to Renaissance Place RT (teachers, administrators, or staff) must enter a user name and password before they can access the data and create reports. Parents who are granted access to Renaissance Place RT must also log in with a user name and password before they can access the Parent Report. 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 Testing Password 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 the scores from one testing occasion to the next, 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 describe 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; however, the STAR Reading Spanish test should only be administered to students who know all their syllables (decode in Spanish).
Content and Item Development

Content Specification

STAR Reading Spanish contains multiple-choice vocabulary-in-context items intended to measure kindergarten through grade 6 reading levels. Content development was driven by the test design and test purposes, which are to measure comprehension and general reading achievement of Spanish text and to place students into appropriate groups for instruction.

The STAR Reading Spanish item bank has 840 operational items. The items are intended to measure a student’s comprehension and general reading achievement of Spanish text.

For information regarding the development of STAR Reading Spanish items, see “Item Development Specifications” on page 9. Assessment items, once written, edited, and reviewed, are field tested and calibrated to estimate their Rasch difficulty parameters and how well they fit to the model. Field testing and calibration are conducted in a single step. Fixed-form assessments were used to collect the item response data needed for psychometric evaluation and calibration analysis. Following these analyses, each assessment item—along with both traditional and 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 to determine whether the item meets all criteria for use in an operational assessment.

The Spanish Vocabulary List

A grade-leveled Spanish Vocabulary List was developed by Renaissance Learning in conjunction with Spanish language experts to provide the set of words to be used throughout item development. Development of the list included four phases:

1. Research-based compilation and creation
2. External expert review
3. Internal review and assessment
4. Refinement

The Spanish Vocabulary List includes 10,000 grade-leveled Spanish words, approximately one thousand words per grade, kindergarten through ninth grade. The list includes the base form of a word—verb infinitives and masculine singular forms of nouns, adjectives, and adverbs. As a validation step, the Spanish Vocabulary List was compared to the Spanish Frequency List, a major frequency
dictionary of Spanish published by Mark Davies, professor of Corpus Linguistics in the Department of Linguistics and English Language at Brigham Young University (2005). Following this comparison, Dr. Adelina Arellano Osuna, professor on the Faculty of Humanities and Education, Center for Research in Foreign Languages (CILE) at the University of the Andes-Mérida, Venezuela, reviewed the list and provided grade-level placement recommendations for each word. Dr. Arellano Osuna's previous work includes a grade-level analysis of words in textbooks and children’s literature used in Central and South American Spanish-speaking classrooms. Renaissance Learning’s editorial teams in the United States and Mexico conducted an internal review of the list for grade-level appropriateness, bias, and completeness. The list of words and grade-level placements were refined during item development.

Item Development Specifications

Valid item development is contingent upon several interdependent factors. The following section outlines the factors which guide STAR Reading Spanish item development. Item content is comprised of stems and answer choices. Additional, detailed information may be found in the English Language Arts Content Appropriateness Guidelines and Item Development Specifications.

Vocabulary-in-Context Item Specifications

Each of the vocabulary items was written to the following specifications:

1. Each vocabulary-in-context test item consists of a single-context sentence. The wording of the sentence is concise and direct. The sentence contains a blank indicating a missing word. Three or four possible answers are shown beneath the sentence. For questions developed at kindergarten through second-grade reading level, three possible answers are given. Questions at a third-grade reading level and higher have 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. All answer choices are chosen from the Spanish Vocabulary List.

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 grade-specific guidelines.

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. Incorrect answer choices are words at the same test level or one grade below. 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.

7. 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.

8. None of these, none of the above, not given, all of the above, and all of these are not used as distractors.

Table 2: Maximum Sentence Length per Item Grade Level

<table>
<thead>
<tr>
<th>Item Grade Level</th>
<th>Maximum Sentence Length (Including Sentence Blank)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindergarten</td>
<td>&lt; 10 words</td>
</tr>
<tr>
<td>Grade 1</td>
<td>10 words</td>
</tr>
<tr>
<td>Grades 2 and 3</td>
<td>12 words</td>
</tr>
<tr>
<td>Grades 4–6</td>
<td>14 words</td>
</tr>
<tr>
<td>Grades 7–13</td>
<td>16 words</td>
</tr>
</tbody>
</table>

Level of Difficulty: Readability

Readability is a primary consideration for level of item difficulty. Readability relates to the overall ease of reading the items. Readability involves the reading level, as well as the layout and visual impact of the stem and answer choices. Readability in STAR Reading Spanish item development accounts for the combined impact, including intensity and density of each part of the item.
The reading level of STAR Reading Spanish items is determined by a combination of factors, including grade-level placement of words on the Spanish Vocabulary list and grade-leveled verb form scope and sequence, sentence difficulty progression, sentence length progression, use of graded target words, and the editorial judgment of both the United States and Mexico-based editorial teams.

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 stimulus components and answer choices. 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. 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 *Diccionario de uso del Español, Segunda Edición*, by María Moliner, is the anchor for spelling. *Esbozo de una Nueva Gramática de la Lengua Española de la Real Academia Española (Comisión de Gramática)* is the reference for grammar, mechanics, and usage.
Item and Scale Calibration

Background

To support a computerized adaptive test of Spanish language reading spanning a number of different grade levels, one major requirement is a bank of test items calibrated as to difficulty, using a vertical, or developmental, scale. Earlier, the development of the test items was described. This chapter describes the research that was conducted in order to calibrate the difficulty of those items and to select some of them for use in the adaptive test.

The research took place in two phases: a pilot phase and the calibration phase. The pilot phase was conducted in the spring of 2011, when 138 new test items were assembled into three forms and administered as fixed-form assessments to tryout samples of students. A total of 1,330 Spanish-speaking students in grades 1 through 5 participated in the pilot study. Each test form included a number of items that were common to all three forms; these were used to link the difficulty scale both within and across the five grades. The analysis of results established that a vertical scale was reasonable for this content and population of students, and provided provisional calibrated difficulty values (Rasch difficulty parameters) for the pilot study test items; 128 of those items were selected to serve as “anchor items” for use in the calibration study.

The calibration phase was much larger, both in numbers of test items and students, and is described in more detail below.

Calibration of STAR Reading Spanish Items

The calibration phase took place in December 2011 and January 2012. A total of 12,839 Spanish-speaking students from volunteering districts took the Spanish-language reading tests. Table 3 shows the summary of the ethnicities of the students who participated in the calibration study. The majority of the students were classified as Hispanic.

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>N</th>
<th>Ethnicity</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian or Alaskan Native</td>
<td>67</td>
<td>Hispanic</td>
<td>4,222</td>
</tr>
<tr>
<td>Asian or Pacific Islander</td>
<td>22</td>
<td>White</td>
<td>136</td>
</tr>
<tr>
<td>Black</td>
<td>18</td>
<td>Not Provided/Missing</td>
<td>5,107</td>
</tr>
</tbody>
</table>
Table 4 shows the distribution of the schools that participated in the STAR Reading Spanish calibration study grouped by geographical region. Of the 74 total schools that participated, the majority were from the South (Texas) and the West (California) regions of the United States.

### Table 4: Participating Schools in the STAR Reading Spanish Calibration Study by Geographic Region

<table>
<thead>
<tr>
<th>Geographic Region</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>2</td>
</tr>
<tr>
<td>Northeast</td>
<td>1</td>
</tr>
<tr>
<td>South</td>
<td>41</td>
</tr>
<tr>
<td>West</td>
<td>30</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>74</strong></td>
</tr>
</tbody>
</table>

The STAR Reading Spanish tests were assembled from a total of 1,825 test items. Of the 1,825 items, 1,699 were new items; 123 were anchor items selected from the first phase; and 3 were practice items, also selected from among the first-phase anchor items. Each of the 1,699 new items was designated as measuring at one of 7 grade levels: kindergarten through grade 6; the 123 anchor items were designated as grade level 1, 2, 3, 4, or 5 items. Table 5 below displays counts of the number and type of items at each grade level.

### Table 5: Numbers of Practice, Anchor, and New Items in Grade Levels K–6 Used in the STAR Reading Spanish 2011–2012 Calibration Study

<table>
<thead>
<tr>
<th>Grade</th>
<th>Practice Items</th>
<th>Anchor Items</th>
<th>New Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>–</td>
<td>–</td>
<td>150</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>28</td>
<td>272</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>25</td>
<td>276</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>23</td>
<td>278</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>22</td>
<td>278</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>25</td>
<td>275</td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>–</td>
<td>170</td>
</tr>
</tbody>
</table>

The calibration tests were administered as adaptive tests using the dynamic calibration features of Renaissance Learning’s STAR application software. Each participating student’s test session included a short practice test, followed by a number of items that were adaptively selected, contingent on performance, as well as a larger number of the new items to be calibrated. Students in grades 1 and
2 responded to a total of 30 items: 3 practice items, 10 scored adaptive items, and 17 adaptively chosen but unscored uncalibrated items. Grade 3 students took a total of 35 items: 3 practice, 11 scored adaptive items, and 21 adaptively chosen unscored items. Students in grades 4 and 5 took 45 items in all: 3 practice, 14 scored adaptive items, and 28 adaptively chosen unscored items. Table 6 displays the design of these adaptive calibration tests, represented as sequences in that grade of:

- practice items (P)
- scored adaptive items (A)
- adaptively chosen unscored new items (C)

Table 6: Adaptive Calibration Test Designs Used for Grades 1–5

<table>
<thead>
<tr>
<th>Grade</th>
<th>Test Length</th>
<th>Item Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>P P P A A A A C C C A C C C C C A C C C C C A A A C C C C C C C C C C</td>
</tr>
<tr>
<td>5</td>
<td>45</td>
<td>P P P A A A A A C C C A C C C C C A A A A A C C C C C A C C C C C C C C C C</td>
</tr>
</tbody>
</table>

The reason for this unusual design was that the pilot study in Phase 1 had shown that there was a weak relationship between student grades and their performance on the Spanish language reading items; some students in the lowest grades read as proficiently as typical students in higher grades; conversely, some students in the higher grades read no better than typical students in the lower grades. For the calibration phase, it was desirable to match students to items appropriate to their ability levels. Because we could not be confident that administering items designated for a specific grade to students at that grade level would accomplish satisfactory matching, an adaptive testing design was employed here. The design called for selecting batches of 3 uncalibrated items at different points in each test, with the grade level of each successive batch chosen contingent on the student’s performance on the scored, adaptive portion of the test up to each point. In this way, students performing well above or below grade level expectations could be matched to appropriate item grade levels.

Because there was no prior information available about students’ reading ability at the outset of the test, each student’s adaptive calibration test began at one of five initial ability levels determined solely by grade. Those ability levels were expressed in the Rasch ability metric, on the vertical scale established in the Phase 1 pilot study. Table 7 lists the starting ability estimate for each grade from 1 to 5.
Calibration students in kindergarten started at the Grade 1 level; those in grades above 5 started at the Grade 5 level.

Table 7: Initial (Starting) Rasch Ability Estimates Used During the Calibration Study for Students in each grades 1–5

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>Starting Ability Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.500</td>
</tr>
<tr>
<td>2</td>
<td>-0.650</td>
</tr>
<tr>
<td>3</td>
<td>0.050</td>
</tr>
<tr>
<td>4</td>
<td>0.700</td>
</tr>
<tr>
<td>5</td>
<td>1.250</td>
</tr>
</tbody>
</table>

Some students took the tests more than once. If the student had taken a test previously during his event, the final ability estimate from that test was used as the starting ability estimate for the student’s later test. Anchor items were administered adaptively. Each Anchor item was selected from the pool based on its match to the current estimate of student ability. A negative offset was added to that current ability estimate. The values of these offsets used to calculate the target Rasch item difficulty during the adaptive calibration tests were as shown in Table 8.

Table 8: Offsets Applied to the Item-by-Item Updated Ability Estimates

<table>
<thead>
<tr>
<th>Anchor Item Sequence</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>-1.73</td>
</tr>
<tr>
<td>Second</td>
<td>-1.10</td>
</tr>
<tr>
<td>Third through end</td>
<td>-0.73</td>
</tr>
</tbody>
</table>

The anchor item selected was the one with the smallest absolute value of the difference between the student’s ability estimate plus the offset and the anchor item’s difficulty calibration (Rasch difficulty parameter). In other words, items were selected with the difficulty parameter closest to the student’s current ability estimate plus the offset.

The pool of appropriate calibration items was determined by the student’s ability estimate, updated as the test progressed, and the grade level of the item as assigned by the item developers. Table 9 displays the ranges of item grade levels assigned to students in each of 7 ranges of estimated ability expressed on the Rasch scale; a student’s estimated Rasch ability varied during the test, depending on performance on the scored adaptive portion at each point.
The research design called for 300 responses per uncalibrated test item, and the recruitment of volunteer participants was targeted toward that goal. Data were actually obtained from 12,839 students, which produced item response counts between 76 and 224 for the new items. There were 612 to 2,973 responses for the items from the pilot study that were used as anchors. Although the target was not reached, there were sufficient responses to warrant proceeding with the item parameter estimation (calibration) analysis.

That analysis proceeded in several steps, and included calibration of both the anchor items and the new items. The final step consisted of concurrent calibration of all items, excluding 28 items that had been rejected in an earlier step on grounds of poor fit to the Rasch model or low correlation to the overall score; 426 students whose responses did not appear to fit the Rasch model were also excluded from the final calibration step. Ultimately, data from 12,413 students were employed in the final step of the calibration analysis; 1,797 of the 1,822 non-practice test items were calibrated. Table 10 displays summary statistics on the Rasch ability scores of the students in the final calibration analysis step, overall and by grade. As expected, students in higher grades scored higher on the calibration study tests. The differences between grades were all statistically significant.

Table 9: Item Grade Levels Used to Select Calibration Items During the Adaptive Calibration Study, Given the Updated Rasch Ability at Any Point in a Student’s Test

<table>
<thead>
<tr>
<th>Student Rasch Ability</th>
<th>Calibration Item Grade Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>–2.0000 or lower</td>
<td>K–2</td>
</tr>
<tr>
<td>–1.999 to –1.0000</td>
<td>K–3</td>
</tr>
<tr>
<td>–0.9999 to –0.3000</td>
<td>K–4</td>
</tr>
<tr>
<td>–0.2999 to 0.4000</td>
<td>1–5</td>
</tr>
<tr>
<td>0.4001 to 1.0000</td>
<td>2–6</td>
</tr>
<tr>
<td>1.0001 to 1.5000</td>
<td>3–6</td>
</tr>
<tr>
<td>above 1.5000</td>
<td>4–6</td>
</tr>
</tbody>
</table>
Note that although the 123 items from the pilot study were used as anchors during the calibration study test administration, the scale being established here was not anchored to the pilot study. In other words, all 1,822 items were calibrated concurrently with no item difficulty parameters set in advance. The test administration engine used the pilot item calibrations as anchors to ensure that students were given items that roughly matched their ability. The item calibrations from the pilot study had no effect on the item difficulty parameters calculated in the calibration phase of the study.

### 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 the majority of 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. Any item that is administered to students of different ability levels (such as different grades) can be expected to have a different p-value at each level.

### 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 is more reliable (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.

**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 logistic (or Rasch) IRT model was selected for the STAR Reading Spanish data both for its simplicity and its ability to accurately model the performance of the STAR Reading Spanish 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 difference between 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 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**

---

**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. Items were eliminated when they met one or more of the following criteria:

- Item-total biserial correlation (item discrimination) was < 0.40
- One or more wrong answer options had an item-total correlation with a positive sign
- 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
Of the initial 1,825 items administered in the STAR Reading Spanish calibration research study, 853 were deemed to be of sufficient quality to be retained for operational use in STAR Reading Spanish.

**Computer-Adaptive Test Design**

The third phase of content specification is determined by the student’s performance during testing. In the conventional paper-and-pencil standardized test, items retained from the item tryout or item calibration study are organized by level; then, each student takes all items within a given test level. Thus, the student is only tested on Spanish reading skills deemed to be appropriate for his or her grade level. 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, a low-performing student’s reading skills may branch to easier items in order to better estimate his or her 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 in Spanish.

Items retained from the STAR Reading Spanish item calibration studies have been organized into a large bank of vocabulary-in-context items, ordered from the easiest to most difficult. During an adaptive 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.

All STAR Reading Spanish tests administer a fixed-length, 25-item, computer-adaptive test. 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 their grade level. The selection of an item that is a bit easier than average minimizes 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 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.
Randomization of items with difficulty values near the student’s adjusted reading ability allows the program to avoid overexposure of test items. All items are dynamically selected from an item bank consisting of the 853 retained vocabulary-in-context 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 minor 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 the STAR Reading Spanish software 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 (currently grades 1–5). Because of this common scale, Scaled Scores are directly comparable with each other, regardless of grade level.

Scale Calibration

The outcome of the item calibration study described above was a sizeable bank of test items suitable for use in the STAR Reading Spanish 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 1–5. 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.
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 is an index of the degree of measurement precision, expressed as a ratio. 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. If the attribute were measured a second time, students’ scores would fluctuate to some degree; that is, scores on the second test would not be perfectly consistent with the same students’ initial scores. The amount of those score fluctuations is an indication of measurement imprecision.

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. 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.

Another way to think of reliability is as an index of 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; one index of reliability 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, a test with 95 percent internal consistency will generally have lower reliability across occasions than it does for a single occasion.

There are a variety of methods of estimating the reliability coefficient of a test. Methods such as Cronbach’s alpha and split-half reliability 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 computer-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 computer-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 conditional standard errors of measurement (CSEM), which provide an index of the degree of error in an individual test score. 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, the CSEM is an estimate of the unreliability of each individual test score. While a reliability coefficient is a single value that applies to the overall test, the magnitude of the CSEM may vary substantially from one person’s test score to another’s.

This chapter presents two different estimates of the internal consistency reliability of STAR Reading Spanish: generic reliability and split-half reliability. Data are not yet available on alternate forms or test-retest reliability, but will be published here when they become available. The reliability discussion here is followed by statistics on the conditional standard error of measurement of STAR Reading Spanish test scores.

The reliability and measurement error presentation describes the reliability coefficients and conditional errors of measurement for the STAR Reading Spanish tests.

**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

\[
\text{reliability} = 1 - \frac{\sigma^2_{\text{error}}}{\sigma^2_{\text{total}}}
\]

where \(\sigma^2_{\text{error}}\) is the variance of the errors of measurement, and \(\sigma^2_{\text{total}}\) 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.

\[
\sigma^2_{\text{error}} = \frac{1}{n} \sum_{i=1}^{n} \text{SEM}^2_i
\]
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 the STAR Reading Spanish research data resulted in the generic reliability estimates shown in the column labeled “Generic” in Table 11 on page 26. 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 would be. These generic reliability coefficients are, therefore, plausible upper-bound estimates of the internal consistency reliability of the STAR Reading Spanish computer-adaptive test.

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) cannot be calculated for adaptive tests. However, an estimate of internal consistency reliability can be calculated using the split-half method. This is discussed in the next section.

Generic reliability estimates are shown in Table 11 on page 26. Results indicated that the overall reliability of the scores was about 0.91. Coefficients ranged from a low of 0.85 in grade 3 to a high of 0.89 in grade 5. These reliability estimates exceeded the rule of thumb that reliability should be a minimum of 0.80 for tests used to make individual pupil decisions.

Overall, these coefficients also compare very favorably with the reliability estimates provided for other published reading tests, which typically contain far more items than the 25-item STAR Reading Spanish. The STAR Reading Spanish test’s high reliability with minimal testing time is a result of careful test item construction and an effective and efficient adaptive-branching procedure.

**Split-Half Reliability**

In classical test theory, before the advent of digital computers automated the calculation of internal consistency reliability measures such as Cronbach’s alpha, approximations such as the split-half method were sometimes used. 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, 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 the first 24 items of the STAR Reading Spanish test; scores based on the odd- and the even-numbered items were calculated. The correlations between the two sets of scores were corrected to a length of 25 items, yielding the split-half reliability estimates displayed in Table 11.

Results indicated that the overall reliability of the scores was about 0.90. The coefficients ranged from a low of 0.85 in grade 3 to a high of 0.90 in grade 1. These reliability estimates are fairly high across grades 1–5 for a test composed of only 25 items, again a result of the measurement efficiency inherent in the adaptive nature of the STAR Reading Spanish test.

### Table 11: Generic and Split-Half Estimates of Internal Consistency Reliability Obtained by Analysis of the Field Research Study Data

<table>
<thead>
<tr>
<th>Grade</th>
<th>N</th>
<th>Generic</th>
<th>Split-Half</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ρxx</td>
<td>ρxx</td>
</tr>
<tr>
<td>1</td>
<td>398</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td>2</td>
<td>496</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>3</td>
<td>483</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>4</td>
<td>181</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>5</td>
<td>145</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>Overall</td>
<td>1,703</td>
<td>0.91</td>
<td>0.90</td>
</tr>
</tbody>
</table>

### 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 can be thought of as a measure of how precise a given score is. 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 32 means that
if a student were tested repeatedly, his or her scores would fluctuate within 32 points of his or her first score about 68 percent of the time, and within 64 points (twice the SEM) roughly 95 percent of the time. Since reliability can also be regarded as a measure of precision, there is a direct relationship between the reliability of a test and the standard error of measurement for the scores it produces.

The STAR Reading Spanish test differs 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 printed tests that report the same SEM value for every examinee. Each administration of the test yields a unique SEM 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 SEM will tend to be lower than that of a conventional test, 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 SEMs for the IRT ability estimates are fairly similar for all students.

Table 12 summarizes the distribution of CSEM values for the research sample, overall and by grade level. The overall average SEM across all grades was 32 scaled score units and ranged from a low of 32 in 1st grade to a high of 33 in 5th grade.

Table 12: Estimate of the Standard Error of Measurement of Scaled Scores, Based on Conditional SEMs and Reliability Estimates

<table>
<thead>
<tr>
<th>Grade</th>
<th>Conditional SEM Method</th>
<th></th>
<th>Split-Half Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>398</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>496</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>483</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>181</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>145</td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td>Overall</td>
<td>1,703</td>
<td>32</td>
<td>3</td>
</tr>
</tbody>
</table>
Validity

The key concept often used to judge an instrument’s usefulness is its validity. The validity of a test is the degree to which it assesses what it claims to measure, and is appropriate for the inferences and decisions made based on its scores. Determining the validity of a test involves the use of data and other information both internal and external to the test instrument itself. One touchstone is content validity, which is the relevance of the test questions to the attributes that are supposed to be measured by the test—namely reading comprehension and reading achievement, in the case of the STAR Reading Spanish test. These content validity issues were discussed in detail in “Content and Item Development” on page 8 and were an integral part of the test items that form the basis of STAR Reading Spanish.

Construct validity, which is the overarching criterion for evaluating a test, investigates the extent to which a test measures the attribute, or construct, 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 Spanish 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.

Three kinds of evidence that would support the construct validity of STAR Reading Spanish are the following:

1. On average, scores should improve with the amount of instruction and experience students have had in Spanish language reading. Therefore, we would expect average scores to increase with students’ grade levels and with age as well.
2. Scores should be related to teachers’ perceptions of the level of their students’ reading skills. Therefore, we would expect a moderate to strong degree of correlation between STAR Reading Spanish scores and formal ratings of students’ reading skills made by their teachers.
3. Scores should be related to established independent measures of Spanish language reading comprehension, skills, and levels of achievement. Therefore, we would expect moderate to strong correlations between scores on STAR Reading Spanish and scores on other tests of similar attributes, such as standardized tests of reading in Spanish, including commercially published tests and state accountability tests.

The remainder of this chapter will be dedicated to presentations of evidence of the three kinds just listed.
STAR Reading Spanish Research Study Data

Since Spanish reading ability varies significantly within and across grade levels and improves as a student’s grade placement increases, scores within STAR Reading Spanish should demonstrate these anticipated internal relationships; in fact, they do.

Table 13 lists summary statistics for STAR Reading Spanish Scaled Scores by age, for those students for whom age was recorded, whereas Table 14 lists summary statistics for STAR Reading Spanish by school grade in the STAR Reading Spanish Research Study. The adaptive test scores from the Research Study increased systematically by age and also by grade. The inter-quartile range statistics show that score variability from the median score was the smallest for the 6-year-old students, but most variable for the 11-year-old students. Score variability by grade was only slight with the most variability occurring in grade 1.

Table 13: Median Scaled Score by Age in the STAR Reading Spanish Field Research Study

<table>
<thead>
<tr>
<th>Age</th>
<th>N</th>
<th>Median Scaled Score</th>
<th>Inter-Quartile Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>85</td>
<td>321</td>
<td>84</td>
</tr>
<tr>
<td>7</td>
<td>304</td>
<td>368</td>
<td>126</td>
</tr>
<tr>
<td>8</td>
<td>361</td>
<td>396</td>
<td>109</td>
</tr>
<tr>
<td>9</td>
<td>353</td>
<td>439</td>
<td>107</td>
</tr>
<tr>
<td>10</td>
<td>155</td>
<td>449</td>
<td>136</td>
</tr>
<tr>
<td>11</td>
<td>87</td>
<td>465</td>
<td>140</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>475</td>
<td>116</td>
</tr>
<tr>
<td>Missing</td>
<td>341</td>
<td>401</td>
<td>147</td>
</tr>
</tbody>
</table>

Table 14: Median Scaled Score by Grade in the STAR Reading Spanish Field Research Study

<table>
<thead>
<tr>
<th>Grade</th>
<th>N</th>
<th>Median Scaled Score</th>
<th>Inter-Quartile Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>398</td>
<td>323</td>
<td>109</td>
</tr>
<tr>
<td>2</td>
<td>496</td>
<td>395</td>
<td>107</td>
</tr>
<tr>
<td>3</td>
<td>483</td>
<td>432</td>
<td>101</td>
</tr>
<tr>
<td>4</td>
<td>181</td>
<td>459</td>
<td>101</td>
</tr>
<tr>
<td>5</td>
<td>145</td>
<td>494</td>
<td>106</td>
</tr>
</tbody>
</table>
Relationship of STAR Reading Spanish to Skills Ratings

In order to have a common measure of Spanish reading skills independent of STAR Reading Spanish, Renaissance Learning constructed a fifteen-item checklist for teachers to use during the Research Study to rate their students on a wide range of competencies related to developing Spanish reading skills. The intent of this checklist was to provide teachers with a single, brief instrument they could use to rate any student from first through fifth grade. In this section, we present the skills rating instrument itself, its psychometric properties as observed in the Research Study, and the relationship between student skills ratings on the instrument and their scores on STAR Reading Spanish.

The Rating Instrument

Teachers completed a Skills Rating form to provide independent ratings of students’ Spanish reading skills. The form included words and sentences at each grade level. Teachers marked the words and sentences that students were able to read independently.

A list of fifteen skill-related items was assembled. Each participating teacher was asked to rate his or her STAR Reading Spanish Research Study students on each skill. To facilitate this, an Excel spreadsheet was designed for each participating class. The rating worksheet consisted of a checklist of fifteen skill-related performance tasks, pre-printed with the names of the participating students. The teacher had simply to mark, for each student, any task he/she believed the student could perform. An example of the rating worksheet is included on the next page.
In the table below please identify which of the following tasks each of your students can probably do correctly.

1. amigo, mover, lindo
2. El gran árbol nos da sombra cuando brilla mucho el sol.
3. Toña no puede hablar porque tiene mucha tos.
4. mochila, grueso, volver
5. Todos los compañeros de la clase estaban invitados a la fiesta de Beto.
6. Mara tuvo que ir al mercado a comprar la verdura para la cena.
7. confianza, murciélago, obligar
8. El detective resolvió el misterio de los juguetes que se habían perdido.
9. El grupo de tercero está orgulloso que su colección de dinosaurios de plástico sea tan grande.
10. conveniente, teatro, perseguir
11. El globo fue descendiendo lentamente a medida que se le salía el aire.
12. Esperamos a que las manzanas estuvieran maduras para poder cosecharlas.
13. brevedad, furgoneta, irregular
14. El canario de la jaula ha estado sentado sobre su percha toda la tarde.
15. Para monitorear la temperatura de Karina, su mamá le estuvo poniendo el termómetro cada hora.

Renaissance Learning, Inc. and its subsidiaries maintain high standards of confidentiality with all data acquired for research and development purposes. Renaissance Learning assures you that all school and student data derived from these activities will only be used for research and development purposes that are intended to validate and/or improve design specifications for general product release into the education market. Individual teacher and student names, grades, and ages will be kept strictly confidential; access to this data will be limited to personnel with relevant research and development responsibilities.

<table>
<thead>
<tr>
<th>Grade 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Student No.</td>
<td>Student Name</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>Palma, Jose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Olivera, Gabi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Palma, Maria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Cruz, Polly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Psychometric Properties of the Skills Ratings

The rating worksheet was scored for each student by assigning one point for each performance task marked by the teacher. The range of possible scores was 0 to 15. Teachers completed skills ratings for 900 of the more than 1,700 students in the Research Study. Table 15 lists data about the psychometric properties of the fifteen-item rating scale, overall and by grade, including the correlations between skills ratings and Scaled Scores. The internal consistency reliability of the skills ratings was 0.94, as estimated by coefficient alpha. Further scrutiny indicated that the survey items were too easy, with little item variance, leading to an inflated coefficient alpha.

Table 15: Summary of Statistics for Skills Ratings and STAR Reading Spanish Scaled Scores, and Correlations Between Them, by Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>N</th>
<th>Rating Mean</th>
<th>Rating Standard Deviation</th>
<th>Scaled Score Mean</th>
<th>Scaled Score Standard Deviation</th>
<th>Correlation of Skills Rating and Scaled Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>274</td>
<td>12.5</td>
<td>4.1</td>
<td>316</td>
<td>96</td>
<td>0.416</td>
</tr>
<tr>
<td>2</td>
<td>305</td>
<td>12.6</td>
<td>3.1</td>
<td>387</td>
<td>95</td>
<td>0.343</td>
</tr>
<tr>
<td>3</td>
<td>232</td>
<td>14.1</td>
<td>2.2</td>
<td>428</td>
<td>81</td>
<td>0.267</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>14.2</td>
<td>2.2</td>
<td>482</td>
<td>71</td>
<td>-0.215</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>14.2</td>
<td>2.6</td>
<td>525</td>
<td>85</td>
<td>-0.080</td>
</tr>
<tr>
<td>Overall</td>
<td>900</td>
<td>13.1</td>
<td>3.2</td>
<td>387</td>
<td>106</td>
<td>0.369</td>
</tr>
</tbody>
</table>

Relationship of Scaled Scores to Skills Ratings

As the data in Table 15 show, the mean skills rating increased from the lowest to the highest grade, from 12.5 at grade 1 to 14.2 at grade 5; however, the increase was mainly between grades 2 and 3. Thus, teachers rated 1st and 2nd grade students as possessing fewer than thirteen of the fifteen skills on average. In contrast, the average third to fifth grader was rated as possessing more than fourteen of the fifteen skills. The correlation between the skills ratings and STAR Reading Spanish Scaled Scores was small at every grade level and also negative for grades 4 and 5. The overall correlation was 0.37, indicating a weak relationship between the computer-adaptive STAR Reading Spanish test scores and teachers’ ratings of their students’ reading skills. The support for validity evidence based on the skills rating scale is equivocal. This could be due to poor psychometric properties of STAR Reading Spanish, the 15-item rating scale, or both. Given that STAR Reading Spanish scores increase systematically by grade as expected, while the skills ratings do not, the rating scale is more likely at fault.
Figure 2 displays the relationships of each of the fifteen rating scale items to STAR Reading Spanish Scaled Scores. These relationships were obtained by fitting mathematical models to the response data for each of the fifteen rating items. Each of the curves in the figure is a graphical depiction of the respective model. As the curves show, the proportion of students rated as possessing each of the fifteen rated skills increases with Scaled Score although most of the skills were very easy; teachers rated more than 80 percent of students at the lowest scale scores as possessing all but four of the 15 skills. Thus, only four or five of the skills rating checklist items (items 10, 13, 14, and 15) differentiated well among students at different STAR Reading Spanish score levels. The items are listed in Table 16.

The relative positions of the curves indicate the relative difficulty of the fifteen rated skills; they are listed from easiest to most difficult in Table 16. The first column of Table 16 indicates the relative difficulty of the fifteen rating items, where relative difficulty 1 is the easiest and 15 is most difficult. The middle column lists the fifteen rated skills, along with their respective item numbers on the rating instrument. The right column lists the “mastery point” for each item. Each mastery point is the lowest STAR Reading Spanish Scaled Score at which a student had at least a 75 percent chance of being rated proficient on the item’s skill. For example,
item 10, “conveniente, teatro, perseguir,” has a mastery point of 313, so students with Scaled Scores below 313 had less than a 75% chance of being rated proficient on that specific skill.

<table>
<thead>
<tr>
<th>Relative Difficulty</th>
<th>Rating Scale Item</th>
<th>Mastery Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easiest</td>
<td>(1) amigo, mover, lindo</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>(2) El gran árbol nos da sombra cuando brilla mucho el sol.</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td>(3) Toña no puede hablar porque tiene mucha tos.</td>
<td>182</td>
</tr>
<tr>
<td></td>
<td>(4) mochila, grueso, volver</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>(6) Mara tuvo que ir al mercado a comprar la verdura para la cena.</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>(5) Todos los compañeros de la clase estaban invitados a la fiesta de Beto.</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>(12) Esperamos a que las manzanas estuvieran maduras para poder cosecharlas.</td>
<td>208</td>
</tr>
<tr>
<td></td>
<td>(8) El detective resolvió el misterio de los juguetes que se habían perdido.</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td>(11) El globo fue descendiendo lentamente a medida que se le salía el aire.</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>(7) confianza, murciélago, obligar</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>(9) El grupo de tercerio está orgulloso que su colección de dinosaurios de plástico sea tan grande.</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>(10) conveniente, teatro, perseguir</td>
<td>313</td>
</tr>
<tr>
<td></td>
<td>(15) Para monitorear la temperatura de Karina, su mamá le estuvo poniendo el termómetro cada hora.</td>
<td>359</td>
</tr>
<tr>
<td></td>
<td>(14) El canario de la jaula ha estado sentado sobre su percha toda la tarde.</td>
<td>405</td>
</tr>
<tr>
<td>Most Difficult</td>
<td>(13) brevedad, furgoneta, irregular</td>
<td>435</td>
</tr>
</tbody>
</table>
Relationship of STAR Reading Spanish Scores to Scores on Other Tests of Spanish Reading Achievement

During the STAR Reading Spanish research study conducted in April 2012, schools submitted data on how their students performed on several tests of Spanish reading achievement as well as their students’ STAR Reading Spanish results. These data included test results for more than 400 students from tests such as the Tejas Lee, Aprenda, the Standards-Based Tests in Spanish, and several statewide tests: the California Standards Test (CST) in English and the State of Texas Assessments of Academic Readiness (STAAR) in English. All the scores were obtained from tests administered in the spring of 2012.

Table 17 displays the validity coefficients for grades 1–5. Blanks in Table 17 indicate that there were no scores available for some tests in some grades. District A provided data for all five grades participating in the study; District B data were limited to grades 1, 2, and 3. Because the reported external tests of Spanish reading comprehension are all grade-specific, the correlations are listed by grade and no overall correlations are reported. The within-grade concurrent validity coefficients for grades 1–5 varied by test. With regard to tests of Spanish reading comprehension, the validity coefficients ranged from 0.50 to 0.87 in the Standards-Based Test in Spanish; 0.77 to 0.89 in Aprenda; 0.35 to 0.58 in the Tejas Lee Accuracy; 0.51 to 0.70 in the Tejas Lee Fluency; and 0.08 to 0.55 in the Tejas Lee Overall Reading Comprehension. Only the first graders had Tejas Lee Reading Comprehension scores with a validity coefficient of 0.77. Although these validity coefficients are mostly supportive of the validity evidence of STAR Reading Spanish scores, the process of establishing the validity of test scores is laborious, and it usually takes a significant amount of time. As a result, the validation of the STAR Reading Spanish test is an ongoing activity, with the goal of establishing evidence of the test’s validity for a variety of settings and students. STAR Reading Spanish users who collect relevant data are encouraged to contact Renaissance Learning.

In general, these correlation coefficients reflect positively on the validity of the STAR Reading Spanish test as a tool for measuring comprehension in Spanish Reading. These validity results, combined with the supporting evidence of reliability, provide a quantitative demonstration of how well this instrument of Spanish reading achievement assessment performs.
The validity data presented in this manual includes evidence of STAR Reading Spanish’s concurrent and construct validity. Table 17 shows evidence of concurrent validity because the external tests were administered at approximately the same time as STAR Reading Spanish in the spring of 2012. Moreover, strong correlations between STAR Reading Spanish test scores with other similar tests of Spanish reading comprehension are evidence of construct validity because such correlations indicate that the tests arguably measure the same underlying reading ability to a large extent.
Score Definitions

This chapter enumerates all of the scores reported by STAR Reading Spanish, including scaled scores and criterion-referenced scores.

Types of Test Scores

STAR Reading Spanish software provides two different types of test scores that measure student performance in different ways: criterion-referenced scores and scaled scores.

Criterion-Referenced Scores

Criterion-referenced scores describe a student’s performance relative to a standard or to a specific content domain. Such scores may be expressed either as a classification, or on a continuous score scale. One 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. An example of a criterion-referenced score on a continuous scale is the Instructional Reading Level reported for both the English and Spanish STAR Reading tests. The STAR Reading Spanish Instructional Reading Level expresses a student’s test performance as the grade level of vocabulary words the student would be expected to recognize and comprehend a specific percent of the time.

Scaled Scores (SS)

STAR Reading Spanish software 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 it is necessary to convert all 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 location on the Rasch ability scale, based on the difficulty of the items administered and the pattern of right and wrong answers. Second, the Rasch ability scores are converted to STAR Reading Spanish Scaled Scores, using a conversion table. STAR Reading Spanish Scaled Scores are reported on a single scale that is independent of grade levels; this is sometimes referred to as a “vertical” or “developmental” scale. STAR Reading Spanish Scaled Scores currently range from 0–790; this range may be extended in the future, when items at higher grade levels are added to the item bank, and field research is extended to include middle and high school students.
Spanish Instructional Reading Level (IRL)

The Instructional Reading Level is a criterion-referenced score that indicates the highest reading level at which the student can most effectively be taught. In other words, IRLs tell you the reading level at which students can recognize words and comprehend written instructional material without 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. Currently, STAR Reading Spanish reports IRLs as Primer (P) and 1.0–6.9.

As a construct, Instructional Reading Levels have existed in the field of reading education for over fifty years. During this time, a variety of assessment instruments have been developed using different measurement criteria that teachers can use to estimate IRL. 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, it uses the student’s Rasch ability scores, in conjunction with the Rasch difficulty parameters of graded vocabulary items, to determine the proportion of items a student can comprehend at each grade level.

Special STAR Reading Spanish Scores

The scaled and IRL scores provided by STAR Reading Spanish software are common measures of reading performance. STAR Reading Spanish software also reports the student’s Zone of Proximal Development.

Spanish Zone of Proximal Development (ZPD)

The Zone of Proximal Development (ZPD) reported by STAR Reading Spanish is the recommended readability (reading difficulty level) range from which students should be selecting books in order to ensure sufficient comprehension and therefore achieve optimal growth in reading skills without experiencing frustration. The Zone of Proximal Development is especially useful for students who use Accelerated Reader, which provides the reading grade levels of over 8,000 Spanish language trade books, magazines, and textbooks. Inspired by the Vygotskian theory, Renaissance Learning developed the ZPD ranges. More information is available in The research foundation for Accelerated Reader goal-setting practices (2006), which is published by Renaissance Learning (http://doc.renlearn.com/KMNet/R001438603GC81D6.pdf).
STAR Reading Spanish in the Classroom

There are numerous ways that STAR Reading Spanish can be used in the classroom, as well as at the school and district level. At the classroom, grade, school, or district level, it can be a useful tool for instructional planning, growth measurement, and program evaluation. At the individual level, it can be used for a variety of purposes, including screening, formative assessment, progress monitoring, and outcomes assessment. This section provides examples of how to use STAR Reading Spanish for many of these purposes.

Formative Assessment

The purpose of formative assessment process is to improve student learning by providing the teacher with instructionally relevant information. STAR Reading Spanish accomplishes this purpose by providing the teacher with valid and reliable information regarding the current reading achievement of students. In many respects, STAR Reading Spanish is comparable to the oral fluency assessment often used for progress monitoring. STAR Reading Spanish is sensitive to slight changes in Spanish reading skills, and it has a high upper range so there is no ceiling effect for most grades. The data generated by STAR Reading Spanish are as useful for instructional planning as are the results of a traditional oral reading fluency assessment.

The Renaissance Learning (2008) Changes to goal-setting and best practices lays out specific recommendations for teachers to improve student learning. These recommendations are based on the findings of large-scale research projects as well as the results of STAR Reading assessments. Among the recommendations are using STAR Reading Spanish to:

- Provide an accurate estimate of students’ current Spanish reading level so
  - Teachers can determine the student’s current level of reading ability in the native language
  - Teachers can match students with appropriate texts for recreational and content-area reading
- Ensure that students are reading more difficult books as their abilities increase
- Identify end-of-year goals for text difficulty
- Help students choose books from different genres that match their interests and challenge their abilities
Measuring Growth

When evaluating or assessing the academic and educational achievement of students, it is important to estimate the amount of growth students obtain within a school year and also across multiple school years. There are many problems inherent in measuring growth from conventional paper and pencil tests within a grade and even more problems associated with measuring growth across multiple grades (see Kolen & Brennan [2004] for more in-depth discussion). STAR Reading Spanish addresses these problems by using a technique called vertical scaling which allows all students’ scores to be placed on the same developmental score scale. This provides comparability within a school year and allows students or cohorts to be followed across multiple school years.

Absolute versus Relative Growth

It is important to distinguish between two types of academic growth (or gains) that may be evidenced in test results: absolute growth and relative growth.

Absolute growth reflects any and all growth that has occurred. For example, as a child begins to read more fluently with practice, we can see absolute growth in the student’s oral reading fluency.

Relative growth reflects only growth that is above and beyond “normal” growth (i.e., beyond typical growth in a reference or norming group). This measure of growth identifies a student’s growth or gains relative to a reference group of students over the same or similar period of time.

The STAR Reading Spanish Growth Report provides you with information about absolute growth.

The construction of STAR Reading Spanish ensures that each student gets approximately equivalent versions of the test at both pretest and posttest administrations, in terms of difficulty to the individual. Thus student growth can be directly measured without any confounding problems related to having seen items at the previous time of measurement. It is important to note that growth is best measured at a group level, such as a classroom or grade level. This is because at the individual student level, there are technical issues of unreliability associated with growth (gain) scores, and measurement error causes fluctuations of individual students’ Scaled Scores that could mask the true amount of growth.

Longitudinal Designs

In longitudinal designs, data is gathered on the same students multiple times over an extended time period. A longitudinal design has at least three time points of
measurement. An example of this approach can be seen in the assessment of students in the fall, winter, and spring quarters of the school year.

The basis for the longitudinal design is to gather ongoing information on student development. This allows for an identification of trends in student achievement along with normal developmental trends with which to compare student growth. Usually, one is interested in how students change over a period of time and finds this change as an indication of instructional and/or intervention efficacy.

Longitudinal designs are very useful as formative evaluations but can also be used in conjunction with summative evaluations. For example, a goal level may be specified for an end-of-the-year evaluation. This would be the summative feature that endeavors to evaluate whether or not the goal was obtained in the time period designated. However, one can incorporate a longitudinal design by more frequently measuring student progress, e.g., at quarterly or monthly intervals. This would allow a teacher to track progress on a monthly basis as the classroom moves towards the stated end-of-year goal. This is also very informative as it provides a signaling system for the teacher if the students begin to fall behind or are not progressing at an expected pace.

Growth Estimates

One important aspect of measuring growth is to have a standard by which to evaluate it. For instance, if someone told you a student gained 25 Scaled Score points in a year, how would you be able to evaluate that information and make a judgment about how well the student is developing? It would be almost impossible without a frame of reference to evaluate the extent to which the student profited from instruction. Therefore, it is important to have some way of interpreting the test score growth a student exhibits. One useful method of doing this would be to relate a student’s growth to an estimate of what would be normal growth for a similar student.

With an estimate of average growth for a student based on growth estimates of similar students, one would then be able to make statements as to whether or not a student made the growth expected within the specific time frame. For instance, many schools and districts use STAR Reading Spanish to measure students at the beginning, middle, and the end of the school year to evaluate how much the school has contributed to the students’ learning. Other schools and districts use STAR Reading Spanish as a summative assessment towards the end of the school year and then use that to gauge growth by the same time at the next school year. Also, now that schools are subject to state accountability regulations in compliance with the No Child Left Behind Act of 2001 (NCLB), many schools now administer a screening assessment at the beginning of the school year to identify students believed to be at-risk of failing to meet the later Spanish reading
standards, and then administer follow-up tests to monitor the progress of these students throughout the school year. STAR Reading Spanish is highly useful for these screening and progress-monitoring functions, given its efficiency, ease of use, and excellent technical qualities.

STAR Reading Spanish’s vertically scaled test scores (scaled scores) allow student scores to be compared across grades as well as within grades. When comparing the growth of students, it is important to have some idea of how much they should be growing normally to evaluate whether or not a program actually increased the growth of a student. Without an expected growth estimate, teachers and administrators may make invalid inferences about the value of a program simply because of normal maturation over time.

In evaluating growth over time, it is important to take grade levels of students into consideration. Two students at different grade levels who attain the same Scaled Score on STAR Reading Spanish may have dramatically different expected growth scores over the same period of time. For instance, suppose a first grader and a second grader both obtain Scaled Scores of 70 on an assessment taken during April of the same school year. It would be wrong to assume that they both should grow the same amount. In fact, a student scoring 70 at the end of first grade would be expected to obtain a Scaled Score of about 133 by the end of the next school year while the second grader would only be expected to score around 118 by the end of the next school year.

Growth is different for different age groups and also different within an age group depending on where students fall in the distribution of abilities. For instance, take the first grade student who scored 70 at the end of the year. This student was expected to score about 63 Scaled Score units higher by the same time in the following school year. However, a similarly aged student in the first grade who scored 140 at the same time would be expected to have a score around 324 by the same time during the next school year. This student is expected to grow by 184 Scaled Score units. Therefore, a single estimate of growth even within a grade can be highly misleading.

To estimate the normal amount of growth from year to year, one must take into account both the grade level of the student at the time of the initial evaluation and also the performance level of the student.
References


Index

A
Absolute growth, 40
  versus relative growth, 40
Accelerated Reader, 38
Access levels and capabilities, 6
Accuracy of content, 12
Adaptive Branching, 1, 5
Adaptive calibration test, 15
Adaptive-branching procedure, 25
Alternate form reliability, 23
Anchor items, 16
Aprenda, 35
Assessments, fixed-form, 13

B
Balanced, demographically and contextually, 11
Bayesian-modal Item Response Theory (IRT), 22

C
Calibration, 8, 13
  Scale, 22
  test, 15
California Standards Test (CST) in English, 35
Capabilities, 6
Classroom, STAR Reading Spanish in the, 39
Cognitive load, 11
Comprehension measure, 8
Computer-adaptive test, 1, 21, 24
Conditional standard error of measurement (CSEM), 22
Construct validity, 28
Content
  development, 8
  differentiation, 11
  specification, 8
Context clues, 10
Contextual and demographic goals, 11
Conventions, language, 12
Correlation coefficients, 35
Criterion-Referenced Scores, 37
Cronbach’s alpha, 23, 25
CSEM (conditional standard error of measurement), 24

D
Data Encryption, 6
Definitions of scores, 37
Demographic and contextual goals, 11
Desktop computer or laptop, use for testing, 2
Dictionary of Spanish, 9
Difficulty calibration (Rasch difficulty parameter), 16
Difficulty parameters, Rasch, 13

E
Empirical item response functions (EIRF), 20
Errors of measurement, 23
Estimated Spanish Instructional Reading Level (Est. SP IRL), 3
Extended time limits, 4

F
Fixed-form assessments, 13
Formative Assessment, 39

G
Generic reliability, 24
Generic reliability coefficients, 25
Goals, demographic and contextual, 11
Grade levels, 42
Grade-level placement, 9
Growth
  absolute versus relative, 40
  estimates, 41
  measuring, 40
  reading, 4
Growth Report, 40

I
In the Classroom, 39
Individualized tests, 5
Initial anxiety, 21
Instructional Reading Levels (IRL), 1
Internal consistency, 23
Internal consistency reliability, 32
iPad®, use for testing, 2
IRT (Item Response Theory), 19, 24
   analysis, 8
   Bayesian-modal, 22
   difficulty scale parameter, 22
   Maximum-Likelihood, 22
Item
   calibration, 13
   development, 8
   difficulty, 18
   discrimination, 18
Item Development Specifications, 9
Item response function (IRF), 19

K
Kuder-Richardson Formula 20 (KR-20), 25

L
Language conventions, 12
Level of difficulty
   Cognitive Load, Content Differentiation, and
   Presentation, 11
   readability, 10
Longitudinal designs, 40

M
Match students with appropriate texts, 39
Maximum-Likelihood IRT (Item Response Theory), 22
Measuring growth, 40

N
No Child Left Behind Act of 2001 (NCLB), 41

O
Oral fluency assessment, 39
Oral reading fluency, 40
Other Spanish Reading Achievement tests
   Aprenda, 35
   California Standards Test (CST) in English, 35
   Standards-Based Tests in Spanish, 35
   State of Texas Assessments of Academic Readiness
   (STAAR) in English, 35
   Tejas Lee, 35
Predicate
   Testing Password, 6
   Presentation, level of difficulty, 11
   Pretest Instructions, 7
Psychometric properties of skills ratings, 32

P
Participating schools, 14
Password entry, 6
Passwords, 6
Phases
   calibration, 13
   pilot, 13
Pilot phase, 13
Practice session, 2
Preference
   Testing Password, 6
Reliability, 2, 23
   alternate form, 23
   coefficient, 23, 24
   coefficients, generic, 25
   coefficients, traditional, 25
   definition, 23
   generic, 24
   internal consistency, 32
   split-half, 23, 25
   test, 24
   test-retest, 23
Relative growth, 40
Restrict student access, 6

STAR Reading Spanish™
Technical Manual
Index

S
Scale calibration, 13, 22
Scaled Score (SS), 3, 22, 24, 29, 32, 37, 41
to skills ratings, 32
Score definitions, 37
Score variability, 29
Scores, test, 23
Scoring, 22
Screening assessment, 41
Security. See Test Security
Single-context sentence, 9
Skill ratings, 30
psychometric properties of, 32
Spanish Instructional Reading Level (SP IRL), 3, 38
Spanish Vocabulary List, 8, 9, 11
Spanish Zone of Proximal Development (ZPD), 38
Spearman-Brown formula, 26
Split-application model, 5
Split-half Reliability, 25
Split-half reliability, 23
Standard error of measurement, 26
Standard time limits, 4
Standards-Based Tests in Spanish, 35
STAR Reading Spanish in the classroom, 39
State of Texas Assessments of Academic Readiness (STAAR)
in English, 35
Student access, restricting, 6
Summary of validity data, 36
Summative assessment, 41
Summative evaluations, 41

T
Tejas Lee, 35
Test
administration procedures, 7
begins, 3
difficulty, 2
interface, 2
length, 3
monitoring, 6
on a desktop computer or laptop, 2
on an iPad®, 2
reliability, 24
repetition, 4
scores, 23
Test security, 5
access levels and capabilities, 6
data encryption, 6
individualized tests, 5
split-application model, 5
Testing Password Preference model, 5
Test-retest reliability, 23
Text difficulty, 39
time limits
extended, 4
standard, 4
time-out limits, 4
Traditional reliability coefficients, 25
Types of test scores
criterion-referenced, 37
scaled score (SS), 37

U
Uncalibrated test item, 17
User names, 6

V
Validity, 28
coefficients, 35
construct, 28
data summary, 36
definition, 28
Vertical scaling, 40
Vocabulary List, Spanish, 8, 9, 11
Vocabulary-in-context
item, 4, 21
item specifications, 9
test items, 2
Vygotskian theory, 38

Z
ZPD (Zone of Proximal Development), Spanish, 1, 38
About Renaissance Learning

Renaissance Learning is a leading provider of cloud-based assessment and teaching and learning solutions that fit the K12 classroom, raise the level of school performance, and accelerate learning for all. By delivering deep insight into what students know, what they like, and how they learn, Renaissance Learning enables educators to deliver highly differentiated and timely instruction while driving personalized student practice in reading, writing, and math every day.

Renaissance Learning leverages top researchers, educators, content-area experts, data scientists, and technologists within a rigorous development and calibration process to deliver and continuously improve its offerings for subscribers in over one-third of U.S. schools and more than 60 countries around the world.