RENAISSANCE®

Star Assessments™ for Reading Technical Manual

RENAISSANCE
Star
Reading®

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

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

Copyright Notice

Copyright © 2019 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, AR 360 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, Subtext, and UClass, 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.

METAMETRICS®, LEXILE®, and LEXILE® FRAMEWORK are trademarks of MetaMetrics, Inc., and are registered in the United States and abroad. Copyright © 2018 MetaMetrics, Inc. All rights reserved.

11/2019 SRRP

Contents

Introduction 1

Star Reading: Screening and Progress-Monitoring Assessment 1

Star Reading Purpose 1

Design of Star Reading 2

Three Generations of Star Reading Assessments 2

Overarching Design Considerations 3

Improvements Specific to Star Reading Versions 3 and Higher 5

Test Interface 6

Practice Session 6

Adaptive Branching/Test Length 7

Test Length 7

Test Repetition 9

Item Time Limits 10

Test Security 11

Split-Application Model 12

Individualized Tests 12

Data Encryption 12

Access Levels and Capabilities 12

Test Monitoring/Password Entry 13

Final Caveat 13

Test Administration Procedures 13

Content and Item Development 14

Content Specification: Star Reading 14

The Educational Development Laboratory's Core Vocabulary List: ATOS Graded Vocabulary List 17

Content Specification: Star Reading 18

Item Development Specifications: Star Reading 18

Vocabulary-in-Context Item Specifications 18

Authentic Text Passage Item Specifications 19

Reading Skills Item Specifications 22
Adherence to Skills 22
Level of Difficulty: Readability 22
Level of Difficulty: Cognitive Load, Content Differentiation, and Presentation 23
Efficiency in Use of Student Time 24
Balanced Items: Bias and Fairness 24
Accuracy of Content 25
Language Conventions 25
Item Components 25

Item and Scale Calibration 27

Background 27
Calibration of Star Reading Items for Use in Version 2 28
Sample Description 28
Item Presentation 30
Item Difficulty 32
Item Discrimination 32
Item Response Function 32
Rules for Item Retention 34
Scale Calibration and Linking 36
On-line Data Collection for New Item Calibration 38
Computer-Adaptive Test Design 40
Scoring in the Star Reading Tests 41
A New Scale for Reporting Star Reading Test Scores 42

Reliability and Measurement Precision 45

34-Item Star Reading Tests 46
Generic Reliability 46
Split-Half Reliability 48
Alternate Form Reliability 48
Standard Error of Measurement 51
25-Item Star Reading Progress Monitoring Tests 54
Reliability Coefficients 54
Standard Error of Measurement 55

Validity **58**

Content Validity 58

Construct Validity 58

Internal Evidence: Evaluation of Unidimensionality of Star Reading 59

External Evidence: Relationship of Star Reading Scores to Scores on Other Tests of Reading Achievement 62

Relationship of Star Reading Scores to Scores on State Tests of Accountability in Reading 65

Relationship of Star Reading Scores to Scores on Multi-State Consortium Tests in Reading 66

Meta-Analysis of the Star Reading Validity Data 67

Additional Validation Evidence for Star Reading 69

- A Longitudinal Study: Correlations with SAT9 69
- Concurrent Validity: An International Study of Correlations with Reading Tests in England 70
- Construct Validity: Correlations with a Measure of Reading Comprehension 71
- Investigating Oral Reading Fluency and Developing the Estimated Oral Reading Fluency Scale 72
- Cross-Validation Study Results 74

Classification Accuracy of Star Reading 75

- Accuracy for Predicting Proficiency on a State Reading Assessment 75
- Accuracy for Identifying At-Risk Students 76
  - Brief Description of the Current Sample and Procedure 77
- Disaggregated Validity and Classification Data 78
- Evidence of Technical Accuracy for Informing Screening and Progress Monitoring Decisions 79
- Screening 80
- Progress Monitoring 84
  - Additional research on Star Reading as a progress monitoring tool 87

Summary of Star Reading Validity Evidence 87

References 88

Norming **89**

Background 89

- The 2017 Star Reading Norms 89

Sample Characteristics 90

Geographic Region 92
School Size 93
Socioeconomic status as indexed by the percent of school students with free and reduced lunch 93
Test Administration 95
Data Analysis 95
Growth Norms 97
Reference 98

Score Definitions 99

Types of Test Scores 99
Scaled Score (SS) 99
Enterprise Scale Scores 100
Unified Scale Scores 100
Grade Equivalent (GE) 101
Comparing the Star Reading Test with Conventional Tests 102
Estimated Oral Reading Fluency (Est. ORF) 102
Instructional Reading Level (IRL) 103
Special IRL Scores 104
Understanding IRL and GE Scores 104
Percentile Rank (PR) 105
Normal Curve Equivalent (NCE) 106
Student Growth Percentile (SGP) 107
Lexile® Measures 108
Lexile Measures of Students and Books: Measures of Student Reading Achievement and Text Readability 109
Special Star Reading Scores 111
Zone of Proximal Development (ZPD) 111
Grade Placement 111
Indicating the Appropriate Grade Placement 112
Compensating for Incorrect Grade Placements 113

Conversion Tables 114

Appendix A: Estimated Oral Reading Fluency 138

Appendix B: Detailed Evidence of Star Reading Validity 139

Relationship of Star Reading Scores to Scores on Other Tests of Reading Achievement 139

Relationship of Star Reading Scores to Scores on State Tests of Accountability in Reading 159

Relationship of Star Reading Enterprise Scores to Scores on Previous Versions 163

Data from Post-Publication Studies 165

Predictive Validity: Correlations with SAT9 and the California Standards Tests 165

Linking Star and State Assessments: Comparing Student- and School-Level Data 166

Methodology Comparison 166

Student-Level Data 167

School-Level Data 167

Accuracy Comparisons 168

Classification Accuracy and Screening Data: NCRTI 171

Aggregated Classification Accuracy Data 172

Receiver Operating Characteristic (ROC) Curves as defined by NCRTI: 172

Brief Description of the Current Sample and Procedure 173

Aggregated Validity Data 175

Disaggregated Validity and Classification Data 175

References 177

Index 179

Introduction

Star Reading: Screening and Progress-Monitoring Assessment

Since the 2011–2012 school year, two different versions of Star Reading have been available for use in assessing the reading achievement of students in grades K–12. The comprehensive version is a 34-item standards-based adaptive assessment, aligned to state and national curriculum standards, that takes an average of less than 20 minutes. A shorter, 25-item version assesses reading comprehension only, and takes an average of less than 10 minutes, making it a popular choice for progress monitoring in programs such as Response to Intervention. Both versions provide immediate feedback to teachers and administrators on each student's reading development.

Star Reading Purpose

As a periodic progress-monitoring assessment, Star Reading progress monitoring serves three purposes for students with at least 100-word sight vocabulary. First, it provides educators with quick and accurate estimates of reading comprehension using students' instructional reading levels. Second, it assesses reading achievement relative to national norms. Third, it provides the means for tracking growth in a consistent manner longitudinally for all students. This is especially helpful to school- and district-level administrators.

The lengthier Star Reading serves similar purposes, but tests a greater breadth of reading skills appropriate to each grade level. While the Star Reading test provides accurate normed data like traditional norm-referenced tests, it is not intended to be used as a "high-stakes" test. Generally, states are required to use high-stakes assessments to document growth, adequate yearly progress, and mastery of state standards. These high-stakes tests are also used to report end-of-period performance to parents and administrators or to determine eligibility for promotion or placement. Star Reading is not intended for these purposes. Rather, because of the high correlation between the Star Reading test and high-stakes instruments, classroom teachers can use Star Reading scores to fine-tune instruction while there is still time to improve performance before the regular test cycle. At the same time, school- and district-level administrators can use Star Reading to predict performance on high-stakes tests. Furthermore, Star Reading results can easily be disaggregated to identify and address the needs of various groups of students.

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

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

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

Star Reading is similar in many ways to the Star Reading Progress Monitoring version, but with some enhanced features, including additional reports and expanded benchmark management.

Design of Star Reading

Three Generations of Star Reading Assessments

The introduction of the current version of Star Reading in 2011 marked the third generation of Star Reading assessments. The first generation consisted of Star Reading version 1, which was a variable-length adaptive assessment of reading comprehension that employed a single item type: vocabulary-in-context items. Star Reading's original item bank contained 800+ such items. Although it was a breakthrough computer adaptive test, Star Reading 1 was based on traditional test theory.

The second generation consisted of Star Reading versions 2 through 4.4, including the current Star Reading Progress Monitoring version. This second generation differed from the first in three major respects: It replaced traditional test theory with Item Response Theory (IRT) as the psychometric foundation for adaptive item selection and scoring; its test length was fixed at twenty-five items (rather than the variable length of version 1); and its content included a second item type: the original vocabulary in context items were augmented in grades 3–12 by the use of longer, authentic text passages for the last 5 items of each test. The second

generation versions differed from one another primarily in terms of the size of their item banks, which grew to over 2000 items in version 4.4. Like the first generation of Star Reading tests, the second generation continued to measure a single construct: reading comprehension.

The third generation is represented by the current version of Star Reading. This is the first version of Star Reading to be designed as a standards-based test; its items are organized into 5 blueprint domains, 10 skill sets, 36 general skills, and over 470 discrete skills—all designed to align to national and state curriculum standards in reading and language arts, including the Common Core State Standards. Like the second generation of Star Reading tests, the third generation Star uses fixed-length adaptive tests. Its tests are longer than the second generation test—34 items in length—both to facilitate broader standards coverage and to improve measurement precision and reliability.

Overarching Design Considerations

One of the fundamental Star Reading design decisions involved the choice of how to administer the test. The primary advantage of using computer software to administer Star Reading tests is the ability to tailor each student's test based on his or her responses to previous items. Conventional assessments, including paper-and-pencil tests, typically entail fixed test forms: every student must respond to the same items in the same sequence. Using computer-adaptive procedures, it is possible for students to test on items that appropriately match their current level of proficiency. The item selection procedures, termed Adaptive Branching, effectively customize the test for each student's achievement level.

Adaptive Branching offers significant advantages in terms of test reliability, testing time, and student motivation. Reliability improves over fixed-form tests because the test difficulty is adjusted to each individual's performance level; students do not have to fit a "one test fits all" model. Most of the test items that students respond to are at levels of difficulty that closely match their achievement level. Testing time decreases because, unlike in paper-and-pencil tests, there is no need to expose every student to a broad range of material, portions of which are inappropriate because they are either too easy for high achievers or too difficult for those with low current levels of performance. Finally, student motivation improves simply because of these issues—test time is minimized and test content is neither too difficult nor too easy.

Another fundamental Star Reading design decision involved the choice of the content and format of items for the test. Many types of stimulus and response procedures were explored, researched, discussed, and prototyped. These procedures included the traditional reading passage followed by sets of literal or inferential questions, previously published extended selections of text followed by

open-ended questions requiring student-constructed answers, and several cloze-type procedures for passage presentation. While all of these procedures can be used to measure reading comprehension and overall reading achievement, the vocabulary-in-context format was selected as the primary item format for the first generation Star Reading assessments. This decision was made for interrelated reasons of efficiency, breadth of construct coverage, and objectivity and simplicity of scoring.

Four fundamental arguments support the use of the original Star Reading design for obtaining quick and reliable estimates of reading comprehension and reading achievement:

1. 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 on 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.

2. In the course of taking the vocabulary-in-context section of Star Reading tests, students read and respond to a significant amount of text. The Star Reading test typically asks the student to demonstrate comprehension of material that ranges over several grade levels. Students will read, use context clues from, interpret the meaning of, and attempt to answer 20 to 25 cloze sentences across these levels, generally totaling more than 300 words. The student must select the correct word from sets of words that are all at the same reading level, and that at least partially fit the sentence context. Students clearly must demonstrate reading comprehension to correctly respond to these 20 to 25 questions.

3. A child's level of vocabulary development is a major factor—perhaps *the* major factor—in determining his or her ability to comprehend written material. Decades of reading research have consistently demonstrated that a student's level of vocabulary knowledge is the most important single element in determining the child's ability to read with comprehension. Tests of vocabulary knowledge typically correlate better than do any other components of reading with valid assessments of reading comprehension. In fact, vocabulary tests often relate more closely to sound measures of reading comprehension than

various measures of comprehension do to each other. Knowledge of word meaning is simply a fundamental component of reading comprehension.

4. The student's performance on the vocabulary-in-context section is used to determine the initial difficulty level of the subsequent authentic text passage items. Although this section consists of just five items, the accurate entry level and the continuing adaptive selection process mean that all of the authentic text passage items are closely matched to the student's reading ability level. This results in unusually high measurement efficiency.

The current third-generation tests expand the breadth of item formats and content beyond that of the previous versions. Each test consists of 34 items; of these, the first 10 are vocabulary-in-context items, while the last 24 items spiral their content to include standards-based material from all five blueprint domains.

The introduction of the 34-Item Star Reading version does not replace the previous version or make it obsolete. The previous version continues to be available as "Star Reading Progress Monitoring," the familiar 25-item measure of reading comprehension. Star Reading thus gives users a choice between a brief assessment focusing on reading comprehension alone, or a longer, standards-based assessment which assures that a broad range of different reading skills, appropriate to student grade level and performance, are included in each assessment.

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

Improvements Specific to Star Reading Versions 3 and Higher

Versions 3 and 4 are adaptations of version 2 designed specifically for use on a computer with web access. In versions 3 and higher, all management and test administration functions are controlled using a management system which is accessed by means of a computer with web access.

This makes a number of new features possible:

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

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

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

Test Interface

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

- If using the keyboard, students press one of the four number keys (**1**, **2**, **3**, and **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 **Next** to enter the answer.
- On a tablet, students tap their answer choice; then, they tap **Next**.

Practice Session

Star Reading software includes a provision for a brief practice test preceding the test itself. The practice session allows students to get comfortable with the test interface and to make sure that they know how to operate it properly. As soon as a student has answered three practice questions correctly, the program takes

the student into the actual test. As long as they possess the requisite 100-word vocabulary, 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 will halt the testing session and tell the student to ask the teacher for help. It may be that the student cannot read at even the most basic level, or it may be that the student needs help operating the interface, in which case the teacher should help the student through the practice session the next time. Before beginning the next test with the student, the program will recommend that the teacher assist the student during the practice.

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

Adaptive Branching/Test Length

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

In order to minimize student frustration, the first administration of the Star Reading test begins with items that have a difficulty level that is below what a typical student at a given grade can handle—usually one or two grades below grade placement. On the average, about 85 percent of students will be able to answer the first item correctly. Teachers can override this typical value by entering an even lower Estimated Instructional Reading Level for the student. On the second and subsequent administrations, the Star Reading test again begins with items that have a difficulty level lower than the previously demonstrated reading ability. Students generally have an 85 percent chance of answering the first item correctly on second and subsequent tests.

Test Length

Once the testing session is underway, the Star Reading test administers 34 items (the Star Reading Progress Monitoring test administers 25 items) of varying difficulty based on the student's responses; this is sufficient information to obtain a reliable Scaled Score and to determine the student's Instructional Reading Level.

The length of time needed to complete a Star Reading test varies across students.

Table 1 provides an overview of the testing time by grade for the students who took the full-length 34-item version of Star Reading during the 2015–2016 school year. The results of the analysis of test completion time indicate that half or more of students completed the test in less than 19 minutes, depending on grade, and even in the slowest grade (grade 1) 95% of students finished their Star Reading test in less than 33 minutes.

Table 2 provides an overview of the Star Reading Progress Monitoring testing time by grade for the students using data from the 2015–16 school year. For that version of the test, about half of the students at every grade completed the Star Reading Progress Monitoring test in less than 10 minutes, and even in the slowest grade (grade 1) 95 percent of students finished in less than 16 minutes.

Table 1: Average and Percentiles of Total Time to Complete the 34-item Star Reading Assessment During the 2015–2016 School Year

Grade	Sample Size	Time to Complete Test (in Minutes)					
		Mean	Standard Deviation	5th Percentile	50th Percentile	95th Percentile	99th Percentile
1	2,024,950	18.65	7.16	9.07	17.63	32.02	39.30
2	4,171,541	19.04	6.47	9.73	18.32	30.87	37.15
3	4,665,013	18.58	5.19	10.35	18.35	27.53	31.40
4	4,282,030	19.17	5.10	10.88	19.03	27.83	31.47
5	3,952,880	19.01	4.89	11.08	18.87	27.33	30.90
6	2,905,934	18.84	4.76	11.07	18.73	26.88	30.33
7	2,311,051	18.44	4.65	10.82	18.35	26.28	29.67
8	2,156,439	18.35	4.61	10.78	18.27	26.12	29.50
9	955,430	18.23	4.66	10.52	18.18	26.03	29.40
10	728,343	18.03	4.68	10.30	17.97	25.88	29.33
11	467,412	17.90	4.75	10.08	17.85	25.87	29.28
12	283,785	17.80	4.91	9.83	17.72	26.07	29.65

Table 2: Average and Percentiles of Total Time to Complete the 25-item Star Reading Progress Monitoring Assessment During the 2015–2016 School Year

Grade	Sample Size	Time to Complete Test (in Minutes)					
		Mean	Standard Deviation	5th Percentile	50th Percentile	95th Percentile	99th Percentile
1	485,266	9.36	3.18	5.48	8.70	15.48	19.40
2	932,641	8.77	2.72	5.43	8.23	13.88	17.12
3	1,064,773	9.16	2.43	5.75	8.87	13.52	15.82
4	1,007,468	8.81	2.31	5.62	8.52	12.98	15.18
5	862,582	8.48	2.20	5.52	8.17	12.50	14.70
6	496,424	8.30	2.13	5.45	7.98	12.20	14.35
7	294,052	8.04	2.08	5.37	7.72	11.83	13.98
8	244,611	7.91	2.02	5.32	7.57	11.70	13.87
9	56,812	7.96	2.05	5.33	7.62	11.75	13.98
10	41,993	7.90	2.04	5.30	7.55	11.72	14.13
11	28,492	7.88	2.02	5.30	7.53	11.67	13.85
12	18,132	7.93	2.12	5.30	7.52	11.88	14.40

Test Repetition

Star Reading score data can be used for multiple purposes such as screening, placement, planning instruction, benchmarking, and outcomes measurement. The frequency with which the assessment is administered depends on the purpose for assessment and how the data will be used. Renaissance Learning recommends assessing students only as frequently as necessary to get the data needed. Schools that use Star for screening purposes typically administer it two to five times per year. Teachers who want to monitor student progress more closely or use the data for instructional planning may use it more frequently. Star Reading may be administered monthly for progress monitoring purposes, and as often as weekly when needed.

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

Item Time Limits

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

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

Table 3 shows the Star Reading Progress Monitoring version's test time-out limits for individual items. These time limits are based on a student's grade level.

Table 3: Star Reading Progress Monitoring Time-Out Limits

Grade	Question Type	Standard Time Limit (seconds/ item)	Extended Time Limit (seconds/item)
K–2	Practice	60	180
	Test, questions 1–25[a]	60	180
	Skill Test—Practice (Calibration)	60	180
	Skill Test—Test (Calibration)	60	180
3–12	Practice	60	180
	Test, questions 1–20[a]	45	135
	Test, questions 21–25[b]	90	270
	Skill Test—Practice (Calibration)	60	180
	Skill Test—Test (Calibration)	90	270

a. Vocabulary-in-context items.
b. Authentic text/passage comprehension items.

These time-out values are based on latency data obtained during item validation. Very few vocabulary-in-context items at any grade had latencies longer than 30 seconds, and almost none (fewer than 0.3 percent) had latencies of more than 45

seconds. Thus, the time-out limit was set to 45 seconds for most students and increased to 60 seconds for the very young students. Longer time limits were allowed for the lengthier authentic text passages items.

Table 4 shows time limits for the 34-item Star Reading version's test questions:

Table 4: Star Reading Time-Out Limits

Grade	Question Type	Standard Time Limit (seconds/item)	Extended Time Limit (seconds/item)
K–2	Practice	60	180
	Test Section A, questions 1–10[a]	60	180
	Test Section B, questions 11–34[b]	120[c]	270[d]
3–12	Practice	60	180
	Test Section A, questions 1–10[a]	45	135
	Test Section B, questions 11–34[b]	90[e]	270[f]

a. Vocabulary-in-context items.
b. Items from 5 domains in 5 blocks, including some vocabulary-in-context.
c. 60 seconds for vocabulary-in-context items.
d. 180 seconds for vocabulary-in-context items.
e. 45 seconds for vocabulary-in-context items.
f. 135 seconds for vocabulary-in-context items.

At all grades, regardless of the extended time limit setting, when a student has only 15 seconds remaining for a given item, a time-out warning appears, indicating that he or she should make a final selection and move on. Items that time out are counted as incorrect responses *unless* the student has the correct answer selected when the item times out. If the correct answer 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 software includes a number of security features to protect the content of the test and to maintain the confidentiality of the test results.

Split-Application Model

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

Individualized Tests

Using Adaptive Branching, every Star Reading 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 students 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 data or access or change it with other software.

Access Levels and Capabilities

Each user's level of access to a Renaissance program depends on the primary position assigned to that user. Each primary position is part of a user permission group. There are six of these groups: district level administrator, district dashboard owner, district staff, school level administrator, school staff, and teacher. By default, each user permission group is granted a specific set of user permissions; each user permission corresponds to one or more tasks that can be performed in the program. The user permissions for these groups can be changed, and user permissions can be granted or removed on an individual level.

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

The security of the Star Reading data is also protected by each person's user name (which must be unique) and password. User names and passwords identify users,

and the program only allows them access to the data and features that they are allowed based on their primary position and the user permissions that they have been granted. Personnel who log in to Renaissance (teachers, administrators, or staff) must enter a user name and password before they can access the data and create reports. Parents who are granted access to Renaissance must also log in with a user name and password before they can access information about their children. Without an appropriate user name and password, personnel and parents cannot use the Star Reading software.

Test Monitoring/Password Entry

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

Final Caveat

While Star Reading 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's security and the quality and validity of its scores.

Test Administration Procedures

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

The Test Administration Manual included with the Star Reading product describes the standard test orientation procedures that teachers should follow to prepare their students for the Star Reading test. These instructions are intended for use with students of all ages; however, the Star Reading test should only be administered to students who have a reading vocabulary of at least 100 words. The instructions were successfully field-tested with students ranging from grades 1–8. It is important to use these same instructions with all students before they take the Star Reading test.

Content and Item Development

Content Specification: Star Reading

The scale and scope of Star Reading content has steadily grown since content development first started many years ago. Since the original test in 1995, which was exclusively Vocabulary-in-Context, other item types have been added to test additional skills, and new item designs continue to be added as state standards evolve.

Star Reading is based upon the assessment of 36 Blueprint Skills organized within 5 Blueprint Domains of reading (see Table 5), and maps the progressions of reading skills and understandings as they develop in sophistication from kindergarten through grade 12. Each Star item is designed to assess a specific skill within the test blueprint. The test blueprint is structured to provide a consistent assessment experience even as state-specific Renaissance Reading Learning Progressions may change, as well as the set of items associated with the blueprint. The Star Reading test blueprint is largely fixed. Renaissance may alter the blueprint if there are data-driven reasons to make a major change to the content.

For information regarding the development of Star Reading items, see "Item Development Specifications: Star Reading" on page 18. Before inclusion in the Star Reading item bank, all items are reviewed to ensure they meet the content specifications for Star Reading item development. Items that do not meet the specifications are either discarded or revised for recalibration. All new item development adheres to the content specifications and all items have been calibrated using the dynamic calibration method.

The first stage of expanded Star Reading development is to identify the set of skills to be assessed. Multiple resources were consulted to determine the set of skills most appropriate for assessing the reading development of K–12 US students. The resources include but are not limited to:

- Reading Next—A Vision for Action and Research in Middle and High School Literacy: A Report to Carnegie Corporation of New York © 2004 by Carnegie Corporation of New York. https://www.all4ed.org/wp-content/uploads/2006/07/ReadingNext.pdf.
- NCTE Principles of Adolescent Literacy Reform, A Policy Research Brief, Produced by The National Council of Teachers of English, April 2006. http://www.ncte.org/library/NCTEFiles/Resources/Positions/Adol-Lit-Brief.pdf.

- Improving Adolescent Literacy: Effective Classroom and Intervention Practices, August 2008. http://eric.ed.gov/PDFS/ED502398.pdf.
- Reading Framework for the 2009 National Assessment of Education Progress. https://www.nagb.gov/assets/documents/publications/frameworks/reading/2009-reading-framework.pdf.
- Common Core State Standards Initiative (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects.
- Individual state standards from all 50 states.

The development of the skills list included iterative reviews by reading and assessment experts and psychometricians specializing in educational assessment. See Table 5 for the Star Reading Blueprint Skills List. The skills list is organized into five blueprint domains:

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

The second stage of development includes item development and calibration. Assessment items are developed according to established specifications for grade-level appropriateness and then reviewed to ensure the items meet the specifications. Grade-level appropriateness is determined by multiple factors including reading skill, reading level, cognitive load, vocabulary grade level, sentence structure, sentence length, subject matter, and interest level. All writers and editors have content-area expertise and relevant classroom experience and use those qualifications in determining grade-level appropriateness for subject matter and interest level. A strict development process is maintained to ensure quality item development.

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

statistics database. A panel of content reviewers then examines each item within the proper context, to determine whether the item meets all criteria for use in an operational assessment.

Table 5: Star Reading Assessment Organization: Star Reading Blueprint Domains, Skill Sets, and Skills

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

An Example of Star Reading Item Adherence to a Specific Skill within Star Reading Blueprint Structure

Blueprint Domain: Analyzing literary text		
Blueprint Skill Set: Literary Elements		
Blueprint Skill: Identify characters and understand characterization		
Grade-level subskill statements:	2nd grade	Describe major and minor characters and their traits using key details.
	3rd grade	Identify and describe main characters' traits, motives, and feelings. **3rd Grade Star Reading Item** Ajay likes being the youngest child in his family. His two older brothers look after him. Before he goes to sleep, they tell him adventure stories. Ajay always falls asleep before the stories are over. The stories will be continued the next night. How does Ajay feel about his brothers? 1. He wants to get bigger so he can play with them. 2. He likes that they look after him and tell him stories. 3. He wishes their stories didn't keep him awake.
	4th grade	Describe characters, interactions with other characters, and relationship between actions, traits, and motives.

The Educational Development Laboratory's Core Vocabulary List: ATOS Graded Vocabulary List

The original point of reference for the development of Star Reading items was the 1995 updated vocabulary lists that are based on the Educational Development Laboratory's (EDL) *A Revised Core Vocabulary* (1969) of 7,200 words. The EDL vocabulary list is a soundly developed, validated list that is often used by developers of educational instruments to create all types of educational materials and assessments. It categorizes hundreds of vocabulary words according to grade placement, from primer (pre-grade 1) through grade 13 (post-high school). This was exactly the span desired for the Star Reading test.

Beginning with new test items introduced in version 4.3, Star Reading item developers have used ATOS instead of the EDL word list. ATOS is a system for evaluating the reading level of continuous text; it contains over 125,000 words in its graded vocabulary list. This readability formula was developed by Renaissance Learning, Inc., and designed by leading readability experts. ATOS is the first formula to include statistics from actual student book reading.

Content Specification: Star Reading

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

The current item bank for Star Reading contains a total of 6,622 items: 2,112 vocabulary-in-context items, 3,849 reading skills items, and 661 authentic text passage items used exclusively in Star Reading Progress Monitoring.

Item Development Specifications: Star Reading

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

Vocabulary-in-Context Item Specifications

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

1. Each vocabulary-in-context test item consists of a single-context sentence. This sentence contains a blank indicating a missing word. Three or four possible answers are shown beneath the sentence. For questions developed at a kindergarten or first-grade reading level, three possible answers are given. Questions at a second-grade reading level and higher offer four possible answers.

2. To answer the question, the student selects the word from the answer choices that best completes the sentence. The correct answer option is the word that appropriately fits both the semantics and the syntax of the sentence. All of the incorrect answer options either fit the syntax of the sentence or relate to the meaning of something in the sentence. They do not, however, meet both conditions.
3. The answer blanks are generally located near the end of the context sentence to minimize the amount of rereading required.
4. The sentence provides sufficient context clues for students to determine the appropriate answer choice. However, the length of each sentence varies according to the guidelines shown in Table 6.
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 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.

Table 6: Maximum Sentence Length per Item Grade Level

Item Grade Level	Maximum Sentence Length (Including Sentence Blank)
Kindergarten and Grade 1	10 words
Grades 2 and 3	12 words
Grades 4–6	14 words
Grades 7–13	16 words

Authentic Text Passage Item Specifications

Authentic Text items are used exclusively as an element of the Star Reading Progress Monitoring test. Authentic text passage items are passages of extended text administered to students at grade levels 3–13. To support students receiving items at grade levels K–3, some original passages were

written. Authentic text items were developed by identifying authentic texts, extracting appropriate passages, and creating cloze-type questions and answers. Each passage is comprised of content that can stand alone as a unified, coherent text. Items were selected which assess passage-level, not merely sentence-level, understanding. To answer the item correctly, the student needs to have a general understanding of the context and content of the passage, not merely an understanding of the specific content of the sentence.

The first authentic passages in Star Reading were extracted from children's and young adult literature, from nonfiction books, and from newspapers, magazines, and encyclopedias. Passages were selected from combinations of three primary categories for school-age children: popular fiction, classic fiction, and nonfiction. Overall Flesch-Kincaid readability estimates of the source materials were used as initial estimates of grade-level difficulty.

After the grade-level difficulty of a passage was estimated, the passage was searched for occurrences of Educational Development Laboratory (EDL) words at the same grade level difficulty. When an EDL word was found that, if replaced with a blank space, would make the passage a good cloze passage, the passage was extracted for use as an authentic text passage test item. Approximately 600 authentic text passage items were initially developed.

Each of the items in the resulting pool was then rated according to several criteria in order to determine which items were best suited for inclusion in the tryout and calibration. Three educators rated each item on the following criteria:

- Grade-level appropriateness of the text
- Cohesiveness of the passage
- Suitability of the passage for its grade level in terms of vocabulary
- Suitability of the passage for its grade level in terms of content density

To ensure a variety of authentic text passage items on the test, each passage was also placed in one of the following categories, according to Meyer and Rice:

1. Antecedent-consequence—causal relationships are found between sentences.
2. Response—a question-answer or a problem-solving format.
3. Comparison—similarities and differences between sentences are found.
4. Collection—sentences are grouped together based on some common idea or event. This would include a sequence of events.
5. Description—sentences provide information by explanation, in specific attributes of the topic, or elaborating on setting.

Replacement passages and newly created items intended for use in versions 4.3 and later were extracted primarily from Accelerated Reader (AR) books. (Updated content specifications were used for writing the new and replacement Star Reading items in version 4.3.) Target words were selected in advance (based on the average ATOS level of target words within a range of difficulty levels). Texts of AR books, based on those with the fewest quiz requests, were run through a text-analysis tool to find instances of use. This was done to decrease the possibility that students may have already encountered an excerpt.

Consideration was given to include some passages from the public domain. When necessary, original long items were written. In any case, passages excerpted or adapted are attributed in "Item and Scale Calibration" on page 27.

Each of the authentic text passage items is written to the following specifications:

1. Each authentic text passage test item consists of a paragraph. The second half of the paragraph contains a sentence with a blank indicating a missing word. Four possible answers are shown beneath the sentence.
2. To answer the question, the student selects the word from the list of answer choices that best completes the sentence based on the context of the paragraph. The correct answer choice is the word that appropriately fits both the semantics and the syntax of the sentence, and the meaning of the paragraph. All of the incorrect answer choices either fit the syntax of the sentence or relate to the meaning of the paragraph.
3. The paragraph provides sufficient context clues for students to determine the appropriate answer choice. Average sentence length within the paragraphs is 8–16 words depending on the item's grade level. Total passage length ranges from 27–107 words, based on the average reading speed of each grade level, as shown in Table 7.

Table 7: Authentic Text Passage Length

Grade	Average Reading Speed (Words/Minute)	Passage Length (Approximate Number of Words)
1	80	30
2	115	40
3	138	55
4	158	70
5–6	173, 185	80
7–9	195, 204, 214	90
10–12	224, 237, 250	100

4. Answer choices for authentic text passage items are EDL Core Vocabulary or ATOS words selected from vocabulary levels at or below that of the correct response. The correct answer for a passage is a word at the targeted level of the item. Incorrect answers are words or appropriate synonyms at the same EDL or ATOS vocabulary level or one grade below.

Reading Skills Item Specifications

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

Adherence to Skills

Star Reading assesses more than 600 grade-specific skills within the Renaissance Reading Learning Progression. Item development is skill-specific. Each item in the item bank is developed for and clearly aligned to one skill. An item meets the alignment criteria if the knowledge and skill required to correctly answer the item match the intended knowledge and skill being assessed. Answering an item correctly does not require reading skill knowledge beyond the expected knowledge for the skill being assessed. Star Reading items include only the information and text needed to assess the skill.

Level of Difficulty: Readability

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

The reading level and grade level for individual words are determined by ATOS. Item stems and answer choices present several challenges to accurately determining reading level. Items may contain discipline-specific vocabulary that is typically above grade level but may still be appropriate for the item. Examples of this could include *summary, paragraph,* or *organized* and the like. Answer choices may be incomplete sentences for which it is difficult to get

an accurate reading grade level. These factors are taken into account when determining reading level.

Item stems and answer choices that are complete sentences are written for the intended grade level of the item. The words in answer choices and stems that are not complete sentences are within the designated grade-level range. Reading comprehension is not complicated by unnecessarily difficult sentence structure and/or vocabulary.

Items and passages are written at grade level. Table 8 indicates the GLE range, item word count range, maximum passage word count range, and sentence length range.

One exception exists for the reading skill *use context clues*. For those items, the target word will be one grade level above the designated grade of the item.

Table 8: Readability Guidelines Table

Grade	GLE Range	Maximum Item Word Count	Sentence Length Range	Number of Words 1 Grade Above (per 100)	Number of Unrecognized Words
K		Less than 30	< 10	0	As a rule, the only unrecognized words will be: names, common derivatives, etc.
1		30	10	0	
2	1.8–2.7	40	Up to 12	0	
3	2.8–3.7	Up to 55	Up to 12	0	
4	3.8–4.7	Up to 70	Up to 14	0	
5	4.8–5.7	Up to 80	Up to 14	In grade 5 and above, only 1 and only when needed.	
6	5.8–6.7	Up to 80	Up to 14	1	
7	6.8–7.7	Up to 90	Up to 16	1	
8	7.8–8.7	Up to 90	Up to 16	1	
9	8.8–9.7	Up to 90	Up to 16	1	
10–12	9.8–10.7	Up to 100	Up to 16	1	

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

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

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

Efficiency in Use of Student Time

Efficiency is evidenced by a good return of information in relation to the amount of time the student spends on the item. The action(s) required of the student are clearly evident. Ideally, the student is able to answer the question without reading the answer choices. Star Reading 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 items adhere to the rules and guidelines in the approved content reference books. *Merriam Webster's 11th Edition* is the reference for pronunciation and spelling. *The Chicago Manual of Style 16th Edition* is the anchor reference for grammar, mechanics, and usage.

Item Components

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

Item stems meet the following criteria with limited exceptions:

- The question is concise, direct, and a complete sentence. The question is written so students can answer it without reading the distractors.
- Generally, completion (blank) stems are not used. If a completion stem is necessary, (such as is the case with vocabulary in context skills) the stem contains enough information for the student to complete the stem without reading the distractors, and the completion blank is as close to the end of the stem as possible.
- The stem does not include verbal or other clues that hint at correct or incorrect distractors.
- The syntax and grammar are straightforward and appropriate for the grade level. Negative construction is avoided.
- The stem does not contain more than one question or part.
- Concepts and information presented in the items are accurate, up-to-

date, and verifiable. This includes but is not limited to dates, references, locations, and events.

Distractors meet the following criteria with limited exceptions:

- All distractors are plausible and reasonable.
- Distractors do not contain clues that hint at correct or incorrect distractors. Incorrect answers are created based on common student mistakes.
- Distractors that are not common mistakes may vary between being close to the correct answer or close to a distractor that is the result of a common mistake.
- Distractors are independent of each other, are approximately the same length, have grammatically parallel structure, and are grammatically consistent with the stem.
- *None of these, none of the above, not given, all of the above,* and *all of these* are not used as distractors.

Item and Scale Calibration

Background

Star Reading was initially published in 1996, and quickly became one of the first applications of computerized adaptive testing (CAT) to educational assessment at the primary and secondary school levels. Unlike other early CAT applications, the initial version of Star Reading was not based on item response theory (IRT). Instead, it was an instance of stratified adaptive testing (Weiss, 1973[1]). The items in its item bank were sorted into grade levels (strata) based on their vocabulary levels. Examinees started the test at the stratum corresponding to their school grade; an algorithm branched them to easier or more difficult levels, contingent on their performance.

IRT was introduced in Version 2 of Star Reading. At that time, hundreds of new test items were developed, and both the new and the original items from Version 1 were calibrated as to difficulty on a vertical scale using the Rasch 1-parameter logistic item response model. Star Reading uses the calibrated Rasch difficulty of the test items as the basis for adaptive item selection. And it uses the Rasch difficulty of the items administered to a student, along with the pattern of right and wrong answers, to calculate a maximum likelihood estimate of the location of the student on the Rasch scale. To provide continuity with the non-IRT score scale of Version 1, equipercentile equating was used to transform the Rasch scores to the original Star Reading score scale.

Version 2's Rasch model-based scale of item difficulty and student ability has continued in use in all subsequent versions of Star Reading. This chapter begins by presenting technical details of the development of that Rasch scale. Later, it will describe improvements that have been made to the method of calibrating the Rasch difficulty of new items. Finally, it will present details of the development of a new scale for reporting Star Reading test scores—the Unified Score Scale, first introduced in the 2017–2018 school year.

1. Weiss, D.J. *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Method Program, 1973. https://eric.ed.gov/?id=ED084301

Calibration of Star Reading Items for Use in Version 2

This section summarizes the psychometric research and development undertaken to prepare the large pool of calibrated reading test questions first used in Star Reading 2, as well as the linkage of Star Reading 2 scores to the original Star Reading 1 score scale. This research took place in two stages: item calibration and score scale calibration. These are described in their respective sections below.

In Star Reading 2 development, a large-scale item calibration program was conducted in the spring of 1998. The Star Reading 2 item calibration study incorporated all of the newly written vocabulary-in-context and authentic text passage items, as well as over 800 vocabulary items in the Star Reading 1 item bank. Two distinct phases comprised the item calibration study. The first phase was the collection of item response data from a multi-level national student sample. The second phase involved the fitting of item response models to the data, and developing a single IRT difficulty scale spanning all levels from grades 1–12.

Sample Description

The data collection phase of the Star Reading 2 calibration study began with a total item pool of over 2100 items. A nationally representative sample of students tested these items. A total of 27,807 students from 247 schools participated in the item calibration study. Table 9 provides the numbers of students in each grade who participated in the study.

Table 9: Numbers of Students Tested by Grade, Star Reading 2 Item Calibration Study—Spring 1998

Grade Level	Number of Students Tested	Grade Level	Number of Students Tested	Grade Level	Number of Students Tested
1	4,037	5	2,167	9	2,030
2	3,848	6	1,868	10	1,896
3	3,422	7	1,126	11	1,326
4	3,322	8	713	12	1,715
				Not Given	337

Table 10 presents descriptive statistics concerning the makeup of the calibration sample. This sample included 13,937 males and 13,626 females (244 student records did not include gender information). As Table 10 As illustrates, the tryout sample approximated the national school population fairly well.

Table 10: Sample Characteristics, Star Reading 2 Calibration Study—Spring 1998 (N = 27,807 Students)

		Students	
		National %	**Sample %**
Geographic Region	Northeast	20%	16%
	Midwest	24%	34%
	Southeast	24%	25%
	West	32%	25%
District Socioeconomic Status	Low: 31–100%	30%	28%
	Average: 15–30%	29%	26%
	High: 0–14%	31%	32%
	Non-Public	10%	14%
School Type & District Enrollment	Public		
	< 200	17%	15%
	200–499	19%	21%
	500–2,000	27%	25%
	> 2,000	28%	24%
	Non-Public	10%	14%

Table 11 provides information about the ethnic composition of the calibration sample. As Table 11 shows, the students participating in the calibration sample closely approximate the national school population.

Table 11: Ethnic Group Participation, Star Reading 2 Calibration Study—Spring 1998 (N = 27,807 Students)

		Students	
		National %	**Sample %**
Ethnic Group	Asian	3%	3%
	Black	15%	13%
	Hispanic	12%	9%
	Native American	1%	1%
	White	59%	63%
	Unclassified	9%	10%

Item Presentation

For the calibration research study, seven levels of test booklets were constructed corresponding to varying grade levels. Because reading ability and vocabulary growth are much more rapid in the lower grades, only one grade was assigned per test level for the first four levels of the test (through grade 4). As grade level increases, there is more variation among both students and school curricula, so a single test can cover more than one grade level. Grades were assigned to test levels after extensive consultation with reading instruction experts as well as considering performance data for items as they functioned in the Star Reading 1 test. Items were assigned to grade levels such that the resulting test forms sampled an appropriate range of reading ability typically represented at or near the targeted grade levels.

Grade levels corresponding to each of the seven test levels are shown in the first two columns of Table 12. Students answered a set number of questions at their current grade level, as well as a number of questions one grade level above and one grade level below their grade level. Anchor items were included to support vertically scaling the test across the seven test levels. Table 12 breaks down the composition of test forms at each test level in terms of types and number of test questions, as well as the number of calibration test forms at each level.

Table 12: Calibration Test Forms Design by Test Level, Star Reading 2 Calibration Study—Spring 1998

Test Level	Grade Levels	Items per Form	Anchor Items per Form	Unique Items per Form	Number of Test Forms
A	1	44	21	23	14
B	2	44	21	23	11
C	3	44	21	23	11
D	4	44	21	23	11
E	5–6	44	21	23	14
F	7–9	44	21	23	14
G	10–12	44	21	23	15

Each of the calibration test forms within a test level consisted of a set of 21 anchor items which were common across all test forms within a test level. Anchor items consisted of items: a) on grade level, b) one grade level above, and c) one grade level below the targeted grade level. The use of anchor items facilitated equating of both test forms and test levels for purposes of data analysis and the development of the overall score scale.

In addition to the anchor items were a set of 23 additional items that were unique to a specific test form (within a level). Items were selected for a specific test level based on Star Reading 1 grade level assignment, EDL vocabulary grade designation, or expert judgment. To avoid problems with positioning effects resulting from the placement of items within each test booklet form, items were shuffled within each test form. This created two variations of each test form such that items appeared in different sequential positions within each "shuffled" test form. Since the final items would be administered as part of a computer-adaptive test, it was important to remove any effects of item positioning from the calibration data so that each item could be administered at any point during the test.

The number of field test forms constructed for each of the seven test levels is shown in the last column of Calibration Test Forms Design by Test Level, Star Reading 2 Calibration Study—Spring 1998 (varying from 11–15 forms per level). Calibration test forms were spiraled within a classroom such that each student received a test form essentially at random. This design ensured that no more than two or three students in any classroom attempted any particular tryout item. Additionally, it ensured a balance of student ability across the various tryout forms. Typically, 250–300 students at the designated grade level of the test item received a given question on their test.

It is important to note that some performance data already existed for the majority of the questions in the Star Reading 2 calibration study. All of the questions from the Star Reading 1 item bank were included, as were many items that were previously field tested, but were not included in the Star Reading 1 test.

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

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

Item Difficulty

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

Item Discrimination

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

Item Response Function

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

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

an item correctly as a function of the student's ability and the difficulty of the item.

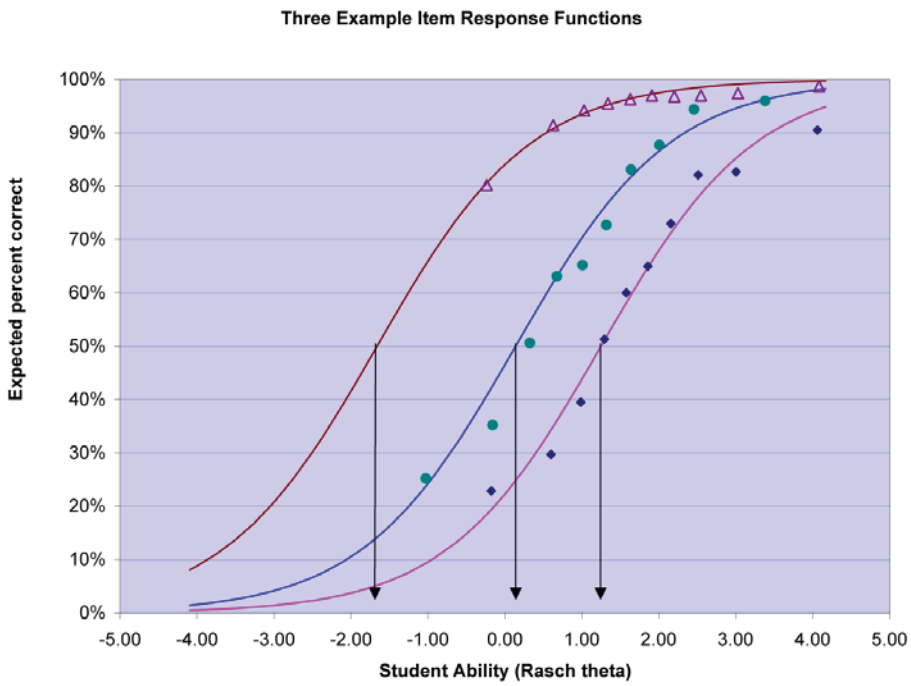
Figure 1 is a plot of three item response functions: one for an easy item, one for a more difficult one, and one for a very difficult item. Each plot is a continuous S-shaped (ogive) curve. The horizontal axis is the scale of student ability, ranging from very low ability (–5.0 on the scale) to very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

In Figure 1, each item's difficulty is the scale point where the expected percent correct is exactly 50. These points are depicted by vertical lines going from the 50 percent point to the corresponding locations on the ability scale. The easiest item has a difficulty scale value of about –1.67; this means that students located at –1.67 on the ability scale have a 50-50 chance of answering that item right. The scale values of the other two items are approximately +0.20 and +1.25, respectively.

Calibration of test items estimates the IRT difficulty parameter for each test item and places all of the item parameters onto a common scale. The difficulty parameter for each item is estimated, along with measures to indicate how well the item conforms to (or "fits") the theoretical expectations of the presumed IRT model.

Also plotted in Figure 1 are "empirical item response functions (EIRF)": the actual percentages of correct responses of groups of students to all three items. Each group is represented as a small triangle, circle, or diamond. Each of those geometric symbols is a plot of the percent correct against the average ability level of the group. Ten groups' data are plotted for each item; the triangular points represent the groups responding to the easiest item. The circles and diamonds, respectively, represent the groups responding to the moderate and to the most difficult item.

Figure 1: Example of Item Statistics Database Presentation of Information



For purposes of the Star Reading 2 calibration research, two different "fit" measures (both unweighted and weighted) were computed. Additionally, if the IRT model is functioning well, then the EIRF points should approximate the (estimated) theoretical IRF. Thus, in addition to the traditional item analysis information, the following IRT-related information was determined for each item administered during the calibration research analyses:

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

Rules for Item Retention

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

information about an item in order to interactively designate items that, in their opinion, did not meet acceptable standards for inclusion in the Star Reading 2 item bank.

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

- Item-total correlation (item discrimination) was < 0.30
- Some other answer option had an item discrimination that was high
- Sample size of students attempting the item was less than 300
- The traditional item difficulty indicated that the item was too difficult or too easy
- The item did not appear to fit the Rasch IRT model

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

Of the initial 2100+ items administered in the Star Reading 2 calibration research study, 1,409 were deemed of sufficient quality to be retained for further analyses. Traditional item-level analyses were conducted again on the reduced data set that excluded the eliminated items. IRT calibration was also performed on the reduced data set and all test forms and levels were equated based on the information provided by the embedded anchor items within each test form. This resulted in placing the IRT item difficulty parameters for all items onto a single scale spanning grades 1–12.

Table 13 summarizes the final analysis information for the test items included in the calibration test forms by test level (A–G). As shown in the table, the item placements in test forms were appropriate: the average percentage of students correctly answering items is relatively constant across test levels. Note, however, that the average scaled difficulty of the items increases across successive levels of the calibration tests, as does the average scaled ability of the students who answered questions at each test level. The median point-biserial correlation, as shown in the table, indicates that the test items were performing well.

Table 13: Calibration Test Item Summary Information by Test Level, Star Reading 2 Calibration Study—Spring 1998

Test Level	Grade Level(s)	Number of Items	Sample Size	Average Percent Correct	Median Percent Correct	Median Point-Biserial	Average Scaled Difficulty	Average Scaled Ability
A	1	343	4,226	67	75	0.56	–3.61	–2.36
B	2	274	3,911	78	88	0.55	–2.35	–0.07
C	3	274	3,468	76	89	0.51	–1.60	0.76
D	4	274	3,340	69	81	0.51	–0.14	1.53
E	5–6	343	4,046	62	73	0.47	1.02	2.14
F	7–9	343	3,875	68	76	0.48	2.65	4.00
G	10–12	366	4,941	60	60	0.37	4.19	4.72

Scale Calibration and Linking

The outcome of the item calibration study described above was a sizable bank of test items suitable for use in the Star Reading 2 test, with an IRT difficulty scale parameter for each item. The item difficulty scale itself was devised such that it spanned a range of item difficulty from grades 1–12. An important feature of Item Response Theory is that the same scale used to characterize the difficulty of the test items is also used to characterize examinees' ability; in fact, IRT models express the probability of a correct response as a function of the difference between the scale values of an item's difficulty and an examinee's ability. The IRT ability/difficulty scale is continuous; values of observed Rasch ability ranged from about –7.0 to +9.0, with the zero value occurring at about the sixth-grade level.

This continuous Rasch score scale is very different from the Scaled Score metric used in Star Reading version 1. Star Reading version 1 scaled scores ranged from 50–1,350, in integer units. The relationship of those scaled scores to the IRT ability scale introduced in Star Reading version 2 was expected to be direct, but not necessarily linear. For continuity between Star Reading 1 and Star Reading 2 scoring, it was desirable to be able to report Star Reading 2 scores on the same scale used in Star Reading 1. To make that possible, a scale linking study was undertaken in conjunction with Star Reading 2 norming. At every grade from 1–12, a portion of the norming sample was asked to take both versions of the Star Reading test: versions 1 and 2. The test score data collected in the course of the linking study were used to link the two scales, providing a conversion table for transforming Star Reading 2 ability scores into equivalent Star Reading 1 Scaled Scores.

From around the country and spanning all 12 grades, 4,589 students participated in the linking study. Linking study participants took both Star Reading 1 and Star Reading 2 tests within a few days of each other. The order in which they took the two test versions was counterbalanced to account for the effects of practice and fatigue. Test score data collected were edited for quality assurance purposes, and 38 cases with anomalous data were eliminated from the linking analyses; the linking was accomplished using data from 4,551 cases. The linking of the two score scales was accomplished by means of an equipercentile equating involving all 4,551 cases, weighted to account for differences in sample sizes across grades. The resulting table of 99 sets of equipercentile equivalent scores was then smoothed using a monotonic spline function, and that function was used to derive a table of Scaled Score equivalents corresponding to the entire range of IRT ability scores observed in the norming study. These Star Reading 2 Scaled Score equivalents range from 0–1400; the same scale has been used for all subsequent Star Reading versions, from version 3 to the present.

Summary statistics of the test scores of the 4,551 cases included in the linking analysis are listed in Table 14. The table lists actual Star Reading 1 Scaled Score means and standard deviations, as well as the same statistics for Star Reading 2 IRT ability estimates and equivalent Scaled Scores calculated using the conversion table from the linking study. Comparing the Star Reading 1 Scaled Score means to the IRT ability score means illustrates how different the two metrics are.

Comparing the Star Reading 1 Scaled Score means to the Star Reading 2 Equivalent Scale Scores in the rightmost two columns of Table 14 illustrates how successful the scale linking was.

Table 14: Summary Statistics of Star Reading 1 and 2 Scores from the Linking Study, by Grade—Spring 1999 (N = 4,551 Students)

		Star Reading 1 Scaled Scores		Star Reading 2 IRT Ability Scores		Star Reading 2 Equivalent Scale Scores	
Grade Level	Sample Size	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	284	216	95	–1.98	1.48	208	109
2	772	339	115	–0.43	1.60	344	148
3	476	419	128	0.33	1.53	419	153
4	554	490	152	0.91	1.51	490	187
5	520	652	176	2.12	1.31	661	213
6	219	785	222	2.98	1.29	823	248
7	702	946	228	3.57	1.18	943	247
8	545	958	285	3.64	1.40	963	276
9	179	967	301	3.51	1.59	942	292
10	81	1,079	292	4.03	1.81	1,047	323
11	156	1,031	310	3.98	1.53	1,024	287
12	63	1,157	299	4.81	1.42	1,169	229
1–12	4,551	656	345	1.73	2.36	658	353

Data from the linking study made it clear that Star Reading 2 software measures ability levels extending beyond the minimum and maximum Star Reading 1 Scaled Scores. In order to retain the superior bandwidth of Star Reading 2 software, extrapolation procedures were used to extend the Scaled Score range below 50 and above 1,350; the range of reported scale scores for Star Reading versions 2 and later is 0 to 1400.

On-line Data Collection for New Item Calibration

As described above, beginning with Star Reading Version 2, item calibration involved administering new items and scale anchoring items to national student samples in printed test booklets. Beginning with Star Reading version 4.3, data needed for item calibration have been collected on-line, by embedding small numbers of uncalibrated items within Star Reading tests. After sufficient numbers of item responses have accumulated, the Rasch difficulty of each new item is estimated by fitting a logistic model to the item response data and the Star Reading Rasch scores of the students' tests. Renaissance Learning calls this overall process "dynamic calibration."

Typically, dynamic calibration is done in batches of several hundred new test items. Each student's test may include between 1 and 5 uncalibrated items. Each item is tagged with a grade level, and is typically administered only to students at that grade level and the next higher grade. The selection of the uncalibrated items to be administered to each student is at random, resulting in nearly equivalent distributions of student ability for each item at a given grade level.

Both traditional and IRT item analyses are conducted of the item response data collected. The traditional analyses yielded proportion correct statistics, as well as biserial and point-biserial correlations between scores on the new items and actual scores on the Star Reading tests. The IRT analyses differed from those used in the calibration of Star Reading 2 items, in that the relationships between scores on each new item and the actual Star Reading scores were used to calibrate the Rasch difficulty parameters.

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

Table 15 summarizes the final analysis information for the 854 new test items introduced in Star Reading Version 4.3, in 2007, by the target grades tagged to each item. Since that time, several thousand more Star Reading items have gone through dynamic calibration; currently the Star Reading operational item bank contains more than 6,500 items.

Table 15: Calibration Test Item Summary Information by Test Item Grade Level, Star Reading 4.3 Calibration Study–Fall 2007

Item Grade Level	Number of Items	Sample Size[a]	Average Percent Correct	Median Percent Correct	Median Point-Biserial	Average Scaled Difficulty	Average Scaled Ability
K	51	230,580	78	78	47	–3.77	–1.65
1	68	238,578	82	82	45	–3.68	–1.23
2	99	460,175	76	76	51	–2.91	–1.06
3	130	693,184	74	78	47	–1.91	–0.23
4	69	543,554	74	78	41	–1.05	0.64
5	44	514,146	70	72	40	–0.14	1.24
6	32	321,855	71	72	38	0.15	1.62
7	42	402,530	60	58	37	1.40	2.07
8	46	317,110	55	53	33	2.10	2.36
9	36	174,906	54	50	33	2.39	2.59
10	56	99,387	51	54	31	2.95	2.91
11	68	62,596	47	43	22	3.50	3.12
12	51	43,343	44	41	18	3.60	3.11
> 12	62	52,359	34	31	11	4.30	3.10

a. Sample size" in this table is the total number of item responses. Each student was presented with 3, 4, or 5 new items, so the sample size substantially exceeds the number of students.

Computer-Adaptive Test Design

In computer-adaptive tests like the Star Reading 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.

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

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

When a student has taken a Star Reading test within the last six months, the difficulty of the first item depends on that student's previous Star Reading test score information. After the administration of the initial item, and after the student has entered an answer, Star Reading 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. Items that have been administered to the same student within the past three-month time period are not available for administration. The large numbers of items available in the item pools, however, ensure that this constraint has negligible impact on the quality of each Star Reading computer-adaptive test.

Scoring in the Star Reading Tests

Following the administration of each Star Reading 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 software uses a proprietary Maximum-Likelihood IRT estimation procedure to avoid any potential of bias in the Scaled Scores.

This approach to scoring enables Star Reading 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 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 (grades K–12). Because of this common scale, Scaled Scores are directly comparable with each other, regardless of grade level. Other scores, such as Percentile Ranks and Grade Equivalents, are derived from the Scaled Scores.

A New Scale for Reporting Star Reading Test Scores

In 2001, five years following the publication of Star Reading Version 1, Renaissance Learning released Star Early Literacy, an assessment of pre-literacy skills that must be developed in order to learn to read. Although the Early Literacy test measures constructs that are different from those assessed in Star Reading, the two assessments are related developmentally, and scores on the two are moderately highly correlated. Over time, many users of Star Reading have also adopted Star Early Literacy; a frequent practice is to transition children from the Early Literacy assessment to Star Reading when they are ready to take the reading assessment. However, the two assessments had very different score scales, making it difficult to recognize the transition point, and impossible to assess growth in cases where Star Early Literacy was used early in the school year, and replaced by Star Reading later in the same year.

What was needed was a common scale that can be used to report scores on both tests. Such a scale, the Unified Score Scale, has been developed, and was introduced into use in the 2017–2018 school year as an optional alternative scale for reporting achievement on both tests.

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

The unified Star Early Learning scale was developed by performing the following steps:

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

Details of these two steps are presented below.

1. The Rasch scale used by Star Early Literacy was linked to the Star Reading Rasch scale.

In this step, a linear transformation of the Star Early Literacy Rasch scale to the Rasch scale used by Star Reading was developed, using a method for linear equating of IRT (item response theory) scales described by Kolen and Brennan (2004, pages 162–165).

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

   Unified Scale Score = INT (42.93 * Star Reading Rasch Score + 958.74)

   where the Star Reading Rasch score has been extended downwards to values as low as –20.00.

   Following are some features and considerations in the development of that scale, called here the "unified scale."

   a. The unified scale's range is from 0 to approximately 1400. Anchor points were chosen such that the 0 point is lower than the Star Reading Rasch scale equivalent of the lowest obtainable SEL scale score, and the lowest obtainable Star Early Literacy (SEL) and Star Reading (SR) scale scores correspond to cardinal numbers on the new scale.

      i. A unified scale score of 100 was set equivalent to an SR Rasch scale score of –20.00.

      ii. The minimum SEL scale score of 300 was set equal to 200 on the unified scale.

      iii. An SR scale score of 0 was set equal to 600 on the unified scale.

   b. The scale uses integer scale scores. New scale scores from 200 to 1400 correspond respectively to the lowest current SEL scale score of 300, and a point slightly higher than the highest current SR scale score of 1400.

   c. The scale is extensible upwards and downwards. Currently, the highest point on the unified scale is just under 1400; but there is no theoretical limit: If SR content were extended beyond the high school reading level, the range of the new scale can be extended upward without limit, as needed. The lowest point is now set at 200—equivalent to the lowest current SEL scale score (300); but the scale can readily be extended downward as low as 0, if a reason arises to do so.

Further details of the transformation of SEL Rasch scores to the SR Rasch scale may be found in the 2018 edition of the *Star Early Literacy Technical Manual.*

Table 16 contains a table of selected Star Reading Rasch ability scores and their equivalents on the Star Reading and Unified Score scales.

Table 16: Some Star Reading Rasch Scores and Their Equivalents on the Star Reading and Unified Score Scales

Minimum Rasch Score	Star Reading Scaled Score	Unified Scale Score
–8.3500	0	600
—6.2845	50	688
–3.1790	100	822
—2.5030	150	851
–1.9030	200	877
–1.2955	250	903
–0.7075	300	928
–0.1805	350	950
0.3390	400	973
0.7600	450	991
1.2450	500	1012
1.6205	550	1028
1.9990	600	1044
2.3240	650	1058
2.5985	700	1070
2.8160	750	1079
3.0090	800	1087
3.2120	850	1096
3.4570	900	1107
3.7435	950	1119
3.9560	1000	1128
4.2120	1050	1139
4.3645	1100	1146
4.5785	1150	1155
4.8280	1200	1166
5.0940	1250	1177
7.5920	1300	1284

Reliability and Measurement Precision

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

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

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

In a computerized adaptive test such as Star Reading, content varies from one administration to another, and it also varies with each student's performance. Another feature of computerized adaptive tests based on Item Response Theory (IRT) is that the degree of measurement error can be expressed for each student's test individually.

The Star Reading 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, 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 test in general, the magnitude of the CSEM may vary substantially from one person's test score to another's.

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

The reliability and measurement error presentation is divided into two sections below: First is a section describing the reliability coefficients and conditional standard errors of measurement for the 34-item Star Reading tests. Second, another brief section presents reliability and measurement error data for the 25-item Star Reading progress monitoring tests. The reliability coefficients and conditional standard errors of measurement are presented for scores expressed on both the Enterprise Star Reading Scale and the newly developed Star Unified Scale.

34-Item Star Reading 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

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

where σ^2_{error} is the variance of the errors of measurement and σ^2_{total} is the variance of test scores. In Star Reading, the variance of the test scores is

easily calculated from Scaled Score data. The variance of the errors of measurement may be estimated from the conditional standard error of measurement (CSEM) statistics that accompany each of the IRT-based test scores, including the Scaled Scores, as depicted below.

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

where the summation is over the squared values of the reported CSEM for students i = 1 to n. In each Star Reading 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 2015–2016 school year data resulted in the generic reliability estimates shown in Table 17 and Table 18 on page 50. Because this method is not susceptible to error variance introduced by repeated testing, multiple occasions, and alternate forms, the resulting estimates of reliability are generally higher than the more conservative alternate forms reliability coefficients. These generic reliability coefficients are, therefore, plausible upper-bound estimates of the internal consistency reliability of the Star Reading computer-adaptive test.

Generic reliability estimates for scores on the unified score scale are shown in Table 17; Table 18 lists the reliability estimates for the older Star Reading "Enterprise" scale scores. Results in Table 17 indicate that the overall reliability of the unified scale scores was about 0.98. Coefficients ranged from a low of 0.94 in grades 4 and 5 to a high of 0.96 in grades 1 and 12. Results based on the Enterprise Scale in Table 18 are slightly lower: the overall reliability of those scale scores was about 0.97; within-grade coefficients ranged from a low of 0.93 in grades 4 to 7 to a high of 0.95 in grades 1, 2,11, and 12.

As both tables show, Star Reading reliability is quite high, grade by grade and overall. Star Reading also demonstrates high test-retest consistency as shown in the rightmost columns of the same tables. Star Reading's technical quality for an interim assessment is on a virtually equal footing with the highest-quality summative assessments in use today.

Split-Half Reliability

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

A split-half reliability coefficient is calculated in three steps. First, the test is divided into two halves, and scores are calculated for each half. Second, the correlation between the two resulting sets of scores is calculated; this correlation is an estimate of the reliability of a half-length test. Third, the resulting reliability value is adjusted, using the Spearman-Brown formula, 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 reliability. These split-half reliability coefficients are independent of the generic reliability approach discussed earlier and more firmly grounded in the item response data. Split-half scores were based on all of the 34 items of the Star Reading tests; scores based on the odd- and the even-numbered items were calculated separately. The correlations between the two sets of scores were corrected to a length of 34 items, yielding the split-half reliability estimates displayed in Table 17 and Table 18 on page 50.

Results indicated that the overall split-half reliability of the Unified scores was 0.98. The coefficients ranged from a low of 0.94 in grades 3 to 8 to a high of 0.96 in grade 12. On the Enterprise Scale, the overall split-half reliability of the Enterprise scores was 0.97. The coefficients ranged from a low of 0.91 in grade 4 to a high of 0.95 in grades 1 and 12. These reliability estimates are quite consistent across grades 1-12, and quite high, again a result of the measurement efficiency inherent in the adaptive nature of the Star Reading test.

Alternate Form Reliability

Another method of evaluating the reliability of a test is to administer the test twice to the same examinees. Next, a reliability coefficient is obtained by calculating the correlation between the two sets of test scores. This is called a test-retest reliability coefficient if the same test was administered both times

and an alternate forms reliability coefficient if different, but parallel, tests were used.

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

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

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

Table 17 and Table 18 include overall and within-grade alternate reliability, along with an indication of the average number of days between testing occasions. The average number of days between testing occasions ranged from 77–117 days.

Results indicated that the overall reliability of the scores on the Unified scale was about 0.94. The alternate form coefficients ranged from a low of 0.81 in grade 12 to a high of 0.87 in grades 3 to 9. Results for the Enterprise scale were similar to those of the Unified Scale with an overall reliability of 0.93; its alternate form coefficients ranged from a low of 0.82 in grade 1 to a high of 0.88 in grades 7–9.

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

Table 17: Reliability Estimates from the Star Reading 2015–2016 Data on the Unified Scale

		Reliability Estimates—Unified Scale				
		Generic	Split-Half	Test-Retest		
Grade	N	ρxx	ρxx	N	ρxx	Average Days between Testing
1	15,583	0.96	0.95	70,000	0.82	78
2	16,489	0.95	0.95	70,000	0.86	78
3	15,349	0.95	0.94	70,000	0.87	77
4	15,641	0.94	0.94	70,000	0.87	81
5	16,131	0.94	0.94	70,000	0.87	83
6	16,221	0.95	0.94	70,000	0.87	92
7	16,381	0.95	0.94	70,000	0.87	99
8	16,573	0.95	0.94	70,000	0.87	100
9	16,277	0.95	0.95	70,000	0.87	113
10	16,575	0.95	0.95	70,000	0.86	114
11	16,326	0.95	0.95	70,000	0.85	117
12	16,098	0.96	0.96	70,000	0.81	114
Overall	193,644	0.98	0.98	840,000	0.94	96

Table 18: Reliability Estimates from the Star Reading 2015–2016 Data on the Enterprise Scale

		Reliability Estimates—Enterprise Scale				
		Generic	Split-Half	Test-Retest		
Grade	N	ρxx	ρxx	N	ρxx	Average Days between Testing
1	15,583	0.95	0.95	70,000	0.82	78
2	16,489	0.95	0.94	70,000	0.86	78
3	15,349	0.94	0.92	70,000	0.87	77
4	15,641	0.93	0.91	70,000	0.87	81
5	16,131	0.93	0.92	70,000	0.87	83
6	16,221	0.93	0.92	70,000	0.87	92
7	16,381	0.93	0.92	70,000	0.88	99
8	16,573	0.94	0.93	70,000	0.88	100
9	16,277	0.94	0.94	70,000	0.88	113
10	16,575	0.94	0.94	70,000	0.87	114
11	16,326	0.95	0.94	70,000	0.86	117
12	16,098	0.95	0.95	70,000	0.84	114
Overall	193,644	0.97	0.97	840,000	0.93	96

Star Reading was designed to be a standards-based assessment, meaning that its item bank measures skills identified by exhaustive analysis of national and state standards in Reading, from grades K–12. The 34-item Star Reading content covers many more skills than Star Reading versions 1 through 4.3, which administered only 25 items.

The increased length of the current version of Star Reading, combined with its increased breadth of skills coverage and enhanced technical quality, was expected to result in improved measurement precision; this showed up as slightly increased reliability, in both internal consistency reliability and alternate form reliability as shown in the tables above. For comparison, see Table 21 on page 54 and Table 22 on page 55.

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 17 means that if a student were tested repeatedly, his or her scores would fluctuate within 17 points of his or her first score about 68 percent of the time, and within 34 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 tests differ from traditional tests in at least two respects with regard to the standard error of measurement. First, Star Reading software computes the SEM for each individual student based on his or her performance, unlike most traditional tests that report the same SEM value for every examinee. Each administration of Star Reading yields a unique "conditional" SEM (CSEM) that reflects the amount of information estimated to be in the specific combination of items that a student received in his or her individual test. Second, because the Star Reading test is adaptive, the CSEM 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 CSEMs for the IRT ability estimates are very similar for all students.

Table 19 and Table 20 contain two different sets of estimates of Star Reading measurement error: conditional standard error of measurement (CSEM) and global standard error of measurement (SEM). Conditional SEM was just described; the estimates of CSEM in Table 19 and Table 20 are the average CSEM values observed for each grade.

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

$$SEM = SQRT(1 - \rho)\ \sigma_X$$

where

SQRT() is the square root operator

ρ is the estimated internal consistency reliability

σ_X is the standard deviation of the observed scores (in this case, Scaled Scores)

Table 19 and Table 20 summarize the distribution of CSEM values for the 2015–2016 data, overall and by grade level. The overall average CSEM on the Unified scale across all grades was 17 scaled score units and ranged from a low of 16 in grades 1–3 to a high of 17 in grades 4–12 (Table 19). The average CSEM based on the Unified scale is similar across all grades. The overall average unified scale score global SEM was 17, the same as the average CSEM. Table 20 shows the average CSEM values on the Enterprise Star Reading scale. Although the adaptive testing process attempts to provide equally precise measurement, regardless of the student's ability level, and the average CSEMs for the IRT ability estimates are very similar for all students, the transformation of the Star Reading IRT ability estimates into equivalent Scaled Enterprise Scores is not linear and the resulting SEMs in the Enterprise Scaled Score metric are less similar.

The overall average CSEM on the Enterprise scale across all grades was 57 scaled score units and ranged from a low of 21 in first grade to a high of 70 in grade 9. Unlike the Unified scale, the Enterprise Scale CSEM values vary by grade and increased with grade until grade 9. The global SEMs for the Enterprise scale scores were higher at each grade, and overall, than the average CSEMs; the overall average SEM was 66. This is attributable to the nonlinear transformation of the Star Reading IRT ability estimates into equivalent Enterprise Scaled Scores. The unified scale, in contrast, is based on a linear transformation of the IRT ability estimates; it eliminates the issues of variable and large CSEM values that are an artifact of the Enterprise Scaled Score nonlinear transformation.

Table 19: Standard Error of Measurement for the 2015–2016 Star Reading Data on the Unified Scale

Grade	Sample Size	Standard Error of Measurement—Unified Scale		
		Conditional		Global
		Average	Standard Deviation	
1	15,583	16	1.1	17
2	16,489	16	1.1	16
3	15,349	16	1.3	17
4	15,641	17	1.3	17
5	16,131	17	1.4	17
6	16,221	17	1.4	17
7	16,381	17	1.5	17
8	16,573	17	1.7	17
9	16,277	17	1.8	17
10	16,575	17	2.2	18
11	16,326	17	2.5	17
12	16,098	17	2.8	17
All	193,644	17	1.8	17

Table 20: Standard Error of Measurement for the 2015–2016 Star Reading Data on the Enterprise Scale

Grade	Sample Size	Standard Error of Measurement—Enterprise Scale		
		Conditional		Global
		Average	Standard Deviation	
1	15,583	21	14.3	27
2	16,489	32	13.0	35
3	15,349	41	15.7	47
4	15,641	49	19.8	58
5	16,131	58	22.9	65
6	16,221	65	24.8	74
7	16,381	69	25.3	78
8	16,573	71	26.3	78
9	16,277	70	27.5	78
10	16,575	69	29.2	76
11	16,326	68	30.1	77
12	16,098	66	31.1	74
All	193,644	57	29.1	66

25-Item Star Reading Progress Monitoring Tests

Star Reading is used for both universal screening and progress monitoring. The 34-item Star Reading test is widely used for universal screening. A shorter version—the 25-item Star Reading progress monitoring test—exists for use in progress monitoring. The following section summarizes the reliability and the standard error of measurement of the progress monitoring version of Star Reading.

Reliability Coefficients

Table 21 and Table 22 show the reliability estimates of the Star Reading progress monitoring test on both the Unified scale and the Enterprise scale.

Table 21: Reliability Estimates from the Star Reading Progress Monitoring Tests on the Unified Scale

	Progress Monitoring Reliability Estimates—Unified Scale						
	Generic		Split-Half		Test-Retest		
Grade	N	ρxx	N	ρxx	N	ρxx	Average Days between Testing
1	100,000	0.95	7,671	0.94	70,000	0.80	83
2	100,000	0.94	8,334	0.93	70,000	0.85	82
3	100,000	0.93	7,946	0.93	70,000	0.85	81
4	100,000	0.93	8,364	0.92	70,000	0.85	85
5	100,000	0.93	8,577	0.92	70,000	0.85	88
6	100,000	0.93	8,656	0.93	70,000	0.85	99
7	100,000	0.94	8,753	0.93	50,000	0.85	112
8	100,000	0.94	8,753	0.93	50,000	0.85	114
9	30,000	0.94	8,626	0.94	10,000	0.86	129
10	30,000	0.95	8,732	0.94	10,000	0.86	133
11	30,000	0.95	8,723	0.94	4,000	0.85	139
12	19,000	0.95	8,640	0.94	4,000	0.82	136
Overall	909,000	0.97	101,775	0.97	548,000	0.93	94

Table 22: Reliability Estimates from the Star Reading Progress Monitoring Tests on the Enterprise Scale

	Progress Monitoring Reliability Estimates—Enterprise Scale						
	Generic		Split-Half		Test-Retest		
Grade	N	ρxx	N	ρxx	N	ρxx	Average Days between Testing
1	100,000	0.93	7,671	0.93	70,000	0.79	83
2	100,000	0.92	8,334	0.91	70,000	0.85	82
3	100,000	0.91	7,946	0.90	70,000	0.85	81
4	100,000	0.90	8,364	0.89	70,000	0.84	85
5	100,000	0.90	8,577	0.89	70,000	0.84	88
6	100,000	0.90	8,656	0.90	70,000	0.84	99
7	100,000	0.91	8,753	0.91	50,000	0.85	112
8	100,000	0.91	8,753	0.91	50,000	0.86	114
9	30,000	0.92	8,626	0.92	10,000	0.87	129
10	30,000	0.93	8,732	0.93	10,000	0.87	133
11	30,000	0.93	8,723	0.93	4,000	0.86	139
12	19,000	0.94	8,640	0.93	4,000	0.85	136
Overall	909,000	0.96	101,775	0.96	548,000	0.93	94

The progress monitoring Star Reading reliability estimates are also quite high and consistent across grades 1–12, for a test composed of only 25 items.

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 progress monitoring tests. The Star Reading progress monitoring test's high reliability with minimal testing time is a result of careful test item construction and an effective and efficient adaptive-branching procedure.

Standard Error of Measurement

Table 23 and Table 24 show the conditional standard error of measurement (CSEM) and the global standard error of measurement (SEM), overall and by grade level.

Table 23: Estimates of Star Reading Progress Monitoring Measurement Precision by Grade and Overall on the Unified Scale

	Progress Monitoring Standard Error of Measurement—Unified Scale				
	Conditional			Global	
Grade	Sample Size	Average	Standard Deviation	Sample Size	SEM
1	100,000	18	2.0	7,671	18
2	100,000	19	1.7	8,334	19
3	100,000	19	1.6	7,946	19
4	100,000	19	1.6	8,364	19
5	100,000	19	1.7	8,577	19
6	100,000	19	1.6	8,656	19
7	100,000	19	1.7	8,753	19
8	100,000	19	2.1	8,753	19
9	30,000	19	2.0	8,626	19
10	30,000	19	2.2	8,732	19
11	30,000	19	2.7	8,723	19
12	19,000	19	3.0	8,640	19
All	909,000	19	1.9	101,775	19

Table 24: Estimates of Star Reading Progress Monitoring Measurement Precision by Grade and Overall on the Enterprise Scale

	Progress Monitoring Standard Error of Measurement—Enterprise Scale				
	Conditional			Global	
Grade	Sample Size	Average	Standard Deviation	Sample Size	SEM
1	100,000	24	16.4	7,671	30
2	100,000	36	14.5	8,334	41
3	100,000	45	16.1	7,946	51
4	100,000	54	20.8	8,364	60
5	100,000	64	25.4	8,577	71
6	100,000	73	28.2	8,656	81
7	100,000	79	30.2	8,753	85
8	100,000	82	31.5	8,753	89
9	30,000	81	33.2	8,626	87
10	30,000	80	34.7	8,732	85
11	30,000	77	36.7	8,723	86
12	19,000	73	38.6	8,640	84
All	909,000	60	32.1	101,775	73

Comparing the estimates of reliability and measurement error of Star Reading (Table 17, Table 18, Table 19, and Table 20) with those of Star Reading progress monitoring (Table 21, Table 22, Table 23, and Table 24) confirms that Star Reading is slightly superior to the shorter Star Reading progress monitoring assessments in terms of reliability and measurement precision.

Validity

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

Content Validity

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

Construct Validity

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

For instance, in a study linking Star Reading Version 1 and the Degrees of Reading Power comprehension assessment, a raw correlation of 0.89 was observed between the two tests. Adjusting that correlation for attenuation

due to unreliability yielded a corrected correlation of 0.96 between the two assessments, indicating that the constructs measured by the different tests are essentially indistinguishable.

Since reading ability varies significantly within and across grade levels and improves as a student's grade placement increases, scores within Star Reading should demonstrate these anticipated internal relationships; in fact, they do. Additionally, scores for Star Reading should correlate highly with other accepted procedures and measures that are used to determine reading achievement and reading comprehension; this is external construct validity. This section deals with both internal and external evidence of the validity of Star Reading as an assessment of reading comprehension and reading skills.

Internal Evidence: Evaluation of Unidimensionality of Star Reading

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

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

Star Reading is an application of item response theory (IRT); each test item's difficulty has been calibrated using the Rasch 1-parameter logistic IRT model. One of the assumptions of the Rasch model is unidimensionality: that a test measures only a single construct such as reading comprehension in the case of Star Reading. To evaluate whether Star reading measures a single construct, factor analyses were conducted. Factor analysis is a statistical technique used to determine the number of dimensions or constructs that a test measures. Both exploratory and confirmatory factor analyses were conducted across grades K to 12.

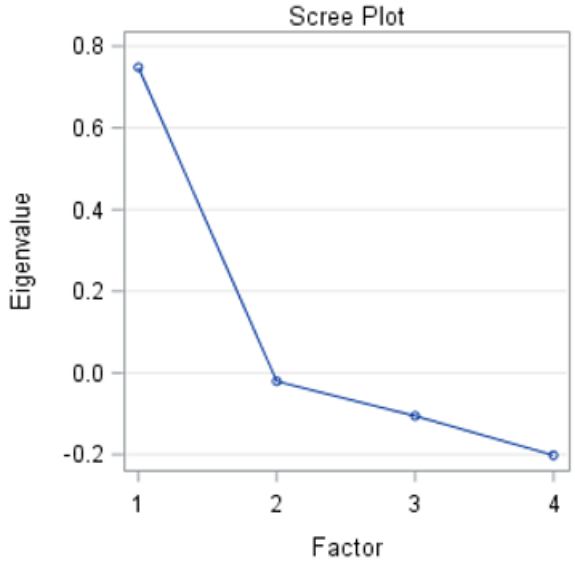
To begin, a large sample of student Star Reading data was assembled. The overall sample consisted of 286,000 student records. That sample was divided into 2 sub-samples. The first sub-sample, consisting of 26,000 cases, was used for exploratory factor analysis; the second sub-sample, 260,000

cases, was reserved for confirmatory factor analyses that followed the initial exploratory analysis.

Within each sub-sample, each student's 34 Star Reading item responses were divided into subsets of items aligned to each of the 5 blueprint domains. Tests administered in grades 4–12 included items from all five domains. Tests given in grades K–3 included items from just 4 domains; no items measuring analyzing argument and evaluating text were administered in these grades. For each student, separate Rasch ability estimates (subtest scores) were calculated from each domain-specific subset of item responses. A Bayesian sequential procedure developed by Owen (1969, 1975) was used for the subtest scoring. The number of items included in each subtest ranged from 2 to 18, following the Star Reading test blueprints, which specify different numbers of items per domain, depending on the student's grade level.

Intercorrelations of the blueprint domain-specific Rasch subtest scores were analyzed using exploratory factor analysis (EFA) to evaluate the number of dimensions/ factors underlying Star Reading. Varimax rotation was used. In each grade, the EFA analyses retained a single dominant underlying dimension based on either the MINEIGEN (eigenvalue greater than 1) or the PROPORTION criterion (proportion of variance explained by the factor), as expected. An example of a scree plot from grade 2 based on the PROPORTION criterion is shown in Figure 2.

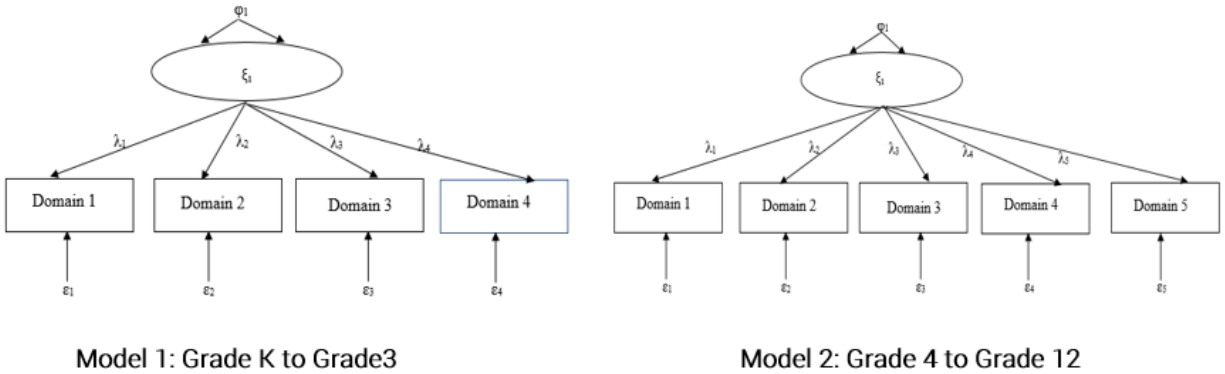
Figure 2: Example Scree Plot from the Grade 2 Exploratory Factor Analysis in Star Reading



Subsequent to the EFA analyses, confirmatory factor analyses (CFA) were also conducted using the subtest scores from the CFA sub-sample. A

separate confirmatory analysis was conducted for each grade. The CFA models tested a single underlying model as shown in Figure 3. Two CFA models were fitted because one of the Star Reading blueprint domains is not tested in grades
K to 3.

Figure 3: Confirmatory Factor Analyses (CFA) in Star Reading



The results of the CFA analyses are summarized in Table 25. As that table indicates, sample sizes ranged from 18,723 to 20,653; because the chi-square (X^2) test is not a reliable test of model fit when sample sizes are large, fit indices are presented. The comparative fit index (CFI), goodness of fit index (GFI), and the normed fit index (NFI) are shown; for these indices, values are either 1 or very close to 1, indicating strong evidence of a single construct/ dimension. In addition, the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) are presented. RMSEA and SRMR values less than 0.08 indicate good fit. Cutoffs for the indices are presented in Hu and Bentler, 1999. Overall, the CFA results strongly support a single underlying construct in Star Reading.

Table 25: Summary of the Goodness-of-Fit of the CFA Models for Star Reading by Grade

Grade	N	X^2	df	CFI	GFI	NFI	RMSEA	SRMR
K	18723	1.4344	2	1.00	1.00	1.00	0.000	0.001
1	19554	7.2567	2	1.00	1.00	1.00	0.012	0.002
2	20608	139.885	2	1.00	1.00	1.00	0.058	0.012
3	19190	125.793	2	1.00	1.00	1.00	0.057	0.013
4	19508	198.639	5	0.99	1.00	0.99	0.045	0.015
5	19958	193.899	5	0.99	1.00	0.99	0.044	0.015
6	20145	225.618	5	0.99	1.00	0.99	0.047	0.015
7	20478	360.502	5	0.99	0.99	0.99	0.059	0.019

Table 25: Summary of the Goodness-of-Fit of the CFA Models for Star Reading by Grade

Grade	N	X^2	df	CFI	GFI	NFI	RMSEA	SRMR
8	20653	405.921	5	0.98	0.99	0.98	0.062	0.021
9	20218	209.196	5	0.99	1.00	0.99	0.045	0.014
10	20619	207.731	5	0.99	1.00	0.99	0.044	0.014
11	20305	243.934	5	0.99	1.00	0.99	0.049	0.014
12	20068	285.21	5	0.99	0.99	0.99	0.053	0.015

The EFA analyses were conducted using the factor procedure while the CFA analysis was conducted using the calis procedure in the SAS 9.4 software (SAS Institute, Cary NC).

External Evidence: Relationship of Star Reading Scores to Scores on Other Tests of Reading Achievement

In an ongoing effort to gather evidence for the validity of Star Reading scores, continual research on score validity has been undertaken. In addition to original validity data gathered at the time of initial development, numerous other studies have investigated the correlations between Star Reading tests and other external measures. In addition to gathering concurrent validity estimates, predictive validity estimates have also been investigated. Concurrent validity was defined for students taking a Star Reading test and external measures within a two-month time period. Predictive validity provides an estimate of the extent to which scores on the Star Reading test predicted scores on criterion measures given at a later point in time, operationally defined as more than two months between the Star test (predictor) and the criterion test. Studies of Star Reading tests' concurrent and predictive correlations with other tests between 1999 and 2013 included the following other tests:

- AIMSweb
- Arkansas Augmented Benchmark Examination (AABE)
- California Achievement Test (CAT)
- Canadian Achievement Test (CAT)
- Colorado Student Assessment Program (CSAP)
- Comprehensive Test of Basic Skills (CTBS)
- Delaware Student Testing Program (DSTP)—Reading

- Dynamic Indicators of Basic Early Literacy Skills (DIBELS)—Oral Reading Fluency
- Florida Comprehensive Assessment Test (FCAT, FCAT 2.0)
- Gates-MacGinitie Reading Test (GMRT)
- Idaho Standards Achievement Test (ISAT)
- Illinois Standards Achievement Test—Reading
- Iowa Test of Basic Skills (ITBS)
- Kansas State Assessment Program (KSAP)
- Kentucky Core Content Test (KCCT)
- Metropolitan Achievement Test (MAT)
- Michigan Educational Assessment Program (MEAP)—English Language Arts and Reading
- Mississippi Curriculum Test (MCT2)
- Missouri Mastery Achievement Test (MMAT)
- New Jersey Assessment of Skills and Knowledge (NJ ASK)
- New York State Assessment Program
- North Carolina End-of-Grade (NCEOG) Test
- Ohio Achievement Assessment (OAA)
- Oklahoma Core Curriculum Test (OCCT)
- South Dakota State Test of Educational Progress (DSTEP)
- Stanford Achievement Test (SAT)
- State of Texas Assessments of Academic Readiness Standards Test (STAAR)
- Tennessee Comprehensive Assessment Program (TCAP)
- TerraNova
- Texas Assessment of Academic Skills (TAAS)
- Transitional Colorado Assessment Program (TCAP)
- West Virginia Educational Standards Test 2 (WESTEST 2)
- Woodcock Reading Mastery (WRM)

- Wisconsin Knowledge and Concepts Examination (WKCE)
- Wide Range Achievement Test 3 (WRAT 3)

26 and Table 27 present summary evidence of concurrent validity collected between 1999 and 2013; between them, these tables summarize some 269 different analyses of concurrent validity with other tests, based on test scores of more than 300 thousand school children. The within-grade average concurrent validity coefficients for grades 1–6 varied from 0.72–0.80, with an overall average of 0.74. The within-grade average concurrent validity for grades 7–12 ranged from 0.65–0.76, with an overall average of 0.72.

Table 28 and Table 29 present summary evidence of predictive validity collected over the same time span: 1999 through 2013. These two tables display summaries of data 300 coefficients of correlation between Star Reading and other measures administered at points in time at least two months later than Star Reading; more than 1.45 million students' test scores are represented in these two tables. Predictive validity coefficients ranged from 0.69–0.72 in grades 1–6, with an average of 0.71. In grades 7–12 the predictive validity coefficients ranged from 0.72–0.87 with an average of 0.80.

In general, these correlation coefficients reflect very well on the validity of the Star Reading test as a tool for placement, achievement and intervention monitoring in Reading. In fact, the correlations are similar in magnitude to the validity coefficients of these measures with each other. These validity results, combined with the supporting evidence of reliability and minimization of SEM estimates for the Star Reading test, provide a quantitative demonstration of how well this innovative instrument in reading achievement assessment performs.

For a compilation of all detailed validation information, see tables of correlations in "Appendix B: Detailed Evidence of Star Reading Validity".

Table 26: Concurrent Validity Data: Star Reading Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6

Summary							
Grade(s)	**All**	**1**	**2**	**3**	**4**	**5**	**6**
Number of students	255,538	1,068	3,629	76,942	66,400	54,173	31,686
Number of coefficients	195	10	18	47	47	41	32
Average validity		0.80	0.73	0.72	0.72	0.74	0.72
Overall average	0.74						

Table 27: Concurrent Validity Data: Star Reading Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 7–12

Summary							
Grade(s)	**All**	**7**	**8**	**9**	**10**	**11**	**12**
Number of students	48,789	25,032	21,134	1,774	755	55	39
Number of coefficients	74	30	29	7	5	2	1
Average validity		0.74	0.73	0.65	0.76	0.70	0.73
Overall average	0.72						

Table 28: Predictive Validity Data: Star Reading Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 1–6

Summary							
Grade(s)	**All**	**1**	**2**	**3**	**4**	**5**	**6**
Number of students	1,227,887	74,887	188,434	313,102	289,571	217,416	144,477
Number of coefficients	194	6	10	49	43	47	39
Average validity		0.69	0.72	0.70	0.71	0.72	0.71
Overall average	0.71						

Table 29: Predictive Validity Data: Star Reading Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 7–12

Summary							
Grade(s)	**All**	**7**	**8**	**9**	**10**	**11**	**12**
Number of students	224,179	111,143	72,537	9,567	21,172	6,653	3,107
Number of coefficients	106	39	41	8	10	6	2
Average validity		0.72	0.73	0.81	0.81	0.87	0.86
Overall average	0.80						

Relationship of Star Reading Scores to Scores on State Tests of Accountability in Reading

The No Child Left Behind (NCLB) Act of 2001 required states to develop and employ their own accountability tests to assess students in ELA/Reading and Math in grades 3 through 8, and one high school grade. Until 2014, most states used their own accountability tests for this purpose. Renaissance Learning was able to obtain accountability test scores for many students who also took Star Reading; in such cases, it was feasible to calculate coefficients of correlation between Star Reading scores and the state test scores. Observed concurrent and predictive validity correlations are summarized

below for the relationship between Star Reading and state accountability test scores for grades 3–8 for tests of both reading and language arts. Table 30 and Table 31 provide summaries from a variety of concurrent and predictive validity coefficients, respectively, for grades 3–8. Numerous state accountability tests have been used in this research.

Table 30: Concurrent Validity Data: Star Reading Correlations (r) with State Accountability Tests, Grades 3–8

Summary							
Grade(s)	**All**	**3**	**4**	**5**	**6**	**7**	**8**
Number of students	11,045	2,329	1,997	2,061	1,471	1,987	1,200
Number of coefficients	61	12	13	11	8	10	7
Average validity		0.72	0.73	0.73	0.71	0.74	0.73
Overall average	0.73						

Table 31: Predictive Validity Data: Star Reading Scaled Scores Predicting Later Performance for Grades 3–8 on Numerous State Accountability Tests

Summary							
Grade(s)	**All**	**3**	**4**	**5**	**6**	**7**	**8**
Number of students	22,018	4,493	2,974	4,086	3,624	3,655	3,186
Number of coefficients	119	24	19	23	17	17	19
Average validity		0.66	0.68	0.70	0.68	0.69	0.70
Overall average	0.68						

For Grades 3 to 8, Star Reading concurrent validity correlations by grade ranged between 0.71 to 0.74 with an overall average validity correlation of 0.71. For Grades 3 to 8, Star Reading predictive validity correlations by grade ranged between 0.66 to 0.70 with an overall average validity correlation of 0.68.

Relationship of Star Reading Scores to Scores on Multi-State Consortium Tests in Reading

In recent years, the National Governors' Association, in collaboration with the Council of Chief State School Officers (CCSSO), developed a proposed set of curriculum standards in English Language Arts and Math, called the Common Core State Standards. Forty-five states voluntarily adopted those standards; subsequently, many states have dropped them, but more than 20 states continue to use them or base their own state standards on them. Two major consortia were formed to develop assessments systems that

embodied those standards: the Smarter Balanced Assessment Consortium (SBAC) and Partnership for Assessment of Readiness for College and Careers (PARCC). SBAC and PARCC end-of-year assessments have been administered in numerous states in place of those states' previous annual accountability assessments. Renaissance Learning was able to obtain SBAC and PARCC scores of many students who had taken Star Reading earlier in the same school years. Table 32 and 33, below, contain coefficients of correlation between Star Reading and the consortium tests.

Table 32: Concurrent Predictive Validity Data: Star Reading Scaled Scores Predicting Later Performance for Grades 3–8 on Smarter Balanced Assessment Consortium Test[a]

Star Reading Predictive and Concurrent Correlations with Smarter Balanced Assessment Scores							
Grade(s)	**All**	**3**	**4**	**5**	**6**	**7**	**8**
Number of students	3,539	709	690	697	567	459	417
Fall Predictive		0.78	0.78	0.76	0.77	0.79	0.80
Winter Predictive		0.78	0.78	0.79	0.78	0.79	0.81
Spring Concurrent		0.79	0.82	0.80	0.70	0.79	0.81

a. Table 32 data are courtesy of the Marysville Washington School District.

Table 33: Concurrent and Predictive Validity Data: Star Reading Scaled Scores Correlations for Grades 3–8 with PARCC Assessment Consortium Test Scores

Star Reading Predictive and Concurrent Correlations with PARCC Assessment Scores							
Grade(s)	**All**	**3**	**4**	**5**	**6**	**7**	**8**
Number of students	22,134	1770	3950	3843	4370	4236	3965
Predictive Concurrent		0.82	0.85	0.82	0.81	0.83	0.80
Concurrent		0.83	0.82	0.78	0.79	0.80	0.77

The average of the concurrent correlations were approximately 0.79 for SBAC and 0.80 for PARCC. The average predictive correlation was 0.78 for the SBAC assessments, and 0.82 for PARCC.

Meta-Analysis of the Star Reading Validity Data

Meta-analysis is a statistical procedure for combining results from different sources or studies. When applied to a set of correlation coefficients that estimate test validity, meta-analysis combines the observed correlations and sample sizes to yield estimates of overall validity. In addition, standard errors and confidence intervals can be computed for overall validity estimates as well as within-grade validity estimates. To conduct a meta-analysis of the Star Reading validity data, 789 correlations reported in the *Star Reading Technical*

Manual were combined and analyzed using a fixed-effects model for meta-analysis (see Hedges and Olkin, 1985, for a methodology description).

The results are displayed in Table 34. The table lists correlations within each grade, as well as results from combining data from all twelve grades. For each set of results, the table gives an estimate of the true validity, a standard error, and the lower and upper limits of a 95 percent confidence interval for the expected validity coefficient. Using the 789 correlation coefficients, the overall estimate of the validity of Star Reading is 0.79, with a standard error of 0.001. The 95 percent confidence interval allows one to conclude that the true validity coefficient for Star Reading is approximately 0.79. The probability of observing the 789 correlations reported in Tables 26–33 if the true validity were zero, would be virtually zero. Because the 789 correlations were obtained with widely different tests, and among students from twelve different grades, these results provide strong support for the validity of Star Reading as a measure of reading skills.

Table 34: Results of the Meta-Analysis of Star Reading Correlations with Other Tests

Grade	Effect Size		95% Confidence Level		Total Correlations	Total N
	Validity Estimate	Standard Error	Lower Limit	Upper Limit		
1	0.70	0.00	0.69	0.70	18	78,022
2	0.78	0.00	0.78	0.78	32	196,114
3	0.79	0.00	0.79	0.79	131	628,336
4	0.79	0.00	0.79	0.79	125	594,712
5	0.79	0.00	0.79	0.79	123	518,411
6	0.79	0.00	0.79	0.79	106	330,475
7	0.79	0.00	0.79	0.79	98	276,218
8	0.79	0.00	0.79	0.79	98	225,704
9	0.78	0.01	0.78	0.79	19	27,952
10	0.82	0.01	0.81	0.82	21	34,913
11	0.74	0.01	0.73	0.74	15	32,798
12	0.86	0.02	0.85	0.87	3	3,146
All Grades	0.79	0.00	0.79	0.79	789	2,946,801

Additional Validation Evidence for Star Reading

This section provides summaries of new validation data along with tables of results. Data from four sources are presented here. They include a predictive validity study, a longitudinal study, a concurrent validity study in England, and a study of Star Reading's construct validity as a measure of reading comprehension.

A Longitudinal Study: Correlations with SAT9

Sadusky and Brem (2002) conducted a study to determine the effects of implementing Reading Renaissance (RR)[1] at a Title I school in the southwest from 1997–2001. This was a retrospective longitudinal study. Incidental to the study, they obtained students' Star Reading posttest scores and SAT9 end-of-year Total Reading scores from each year and calculated correlations between them. Students' test scores were available for multiple years, spanning grades 2–6. Data on gender, ethnic group, and Title I eligibility were also collected.

Table 35 displays the observed correlations for the overall group. Table 36 displays the same correlations, broken out by ethnic group.

Overall correlations by year ranged from 0.66–0.73. Sadusky and Brem concluded that "Star results can serve as a moderately good predictor of SAT9 performance in reading."

Enough Hispanic and white students were identified in the sample to calculate correlations separately for those two groups. Within each ethnic group, the correlations were similar in magnitude, as Table 36 shows. This supports the assertion that Star Reading is valid for multiple student ethnicities.

Table 35: Correlations of the Star Posttest with the SAT9 Total Reading Scores 1998–2002[a]

Year	Grades	N	Correlation
1998	3–6	44	0.66
1999	2–6	234	0.69
2000	2–6	389	0.67
2001	2–6	361	0.73

a. All correlations significant, $p < 0.001$.

1. Reading Renaissance is a supplemental reading program that uses Star Reading and Accelerated Reader.

Table 36: Correlations of the Star Posttest with the SAT9 Total Reading Scores, by Ethnic Group, 1998–2002[a]

Year	Grade	Hispanic		White	
		N	Correlation	N	Correlation
1998	3–6	7 (n.s.)	0.55	35	0.69
1999	2–6	42	0.64	179	0.75
2000	2–6	67	0.74	287	0.71
2001	2–6	76	0.71	255	0.73

a. All correlations significant, $p < 0.001$, unless otherwise noted.

Concurrent Validity: An International Study of Correlations with Reading Tests in England

NFER, the National Foundation for Educational Research, conducted a study of the concurrent validity of both Star Reading and Star Math in 16 schools in England in 2006 (Sewell, Sainsbury, Pyle, Keogh and Styles, 2007). English primary and secondary students in school years 2–9 (equivalent to US grades 1–8) took both Star Reading and one of three age-appropriate forms of the Suffolk Reading Scale 2 (SRS2) in the fall of 2006. Scores on the SRS2 included traditional scores, as well as estimates of the students' Reading Age (RA), a scale that is roughly equivalent to the Grade Equivalent (GE) scores used in the US. Additionally, teachers conducted individual assessments of each student's attainment in terms of curriculum levels, a measure of developmental progress that spans the primary and secondary years in England.

Correlations with all three measures are displayed in Table 37, by grade and overall. As the table indicates, the overall correlation between Star Reading and Suffolk Reading Scaled Scores was 0.91, the correlation with Reading Age was 0.91, and the correlation with teacher assessments was 0.85. Within-form correlations with the SRS ability estimate ranged from 0.78–0.88, with a median correlation of 0.84, and ranged from 0.78–0.90 on Reading Age, with a median of 0.85.

Table 37: Correlations of Star Reading with Scores on the Suffolk Reading Scale and Teacher Assessments in a Study of 16 Schools in England

	Suffolk Reading Scale				Teacher Assessments	
School Years[a]	Test Form	N	SRS Score[b]	Reading Age	N	Assessment Levels
2–3	SRS1A	713	0.84	0.85	n/a	n/a
4–6	SRS2A	1,255	0.88	0.90	n/a	n/a
7–9	SRS3A	926	0.78	0.78	n/a	n/a
Overall		2,694	0.91	0.91	2,324	0.85

a. UK school year values are 1 greater than the corresponding US school grade. Thus, Year 2 corresponds to Grade 1, etc.
b. Correlations with the individual SRS forms were calculated with within-form raw scores. The overall correlation was calculated with a vertical Scaled Score.

Construct Validity: Correlations with a Measure of Reading Comprehension

The Degrees of Reading Power (DRP) test is widely recognized as a measure of reading comprehension. Yoes (1999) conducted an analysis to link the Star Reading Rasch item difficulty scale to the item difficulty scale of DRP. As part of the study, nationwide samples of students in grades 3, 5, 7, and 10 took two tests each (leveled forms of both the DRP and of Star Reading calibration tests). The forms administered were appropriate to each student's grade level. Both tests were administered in paper-and-pencil format. All Star Reading test forms consisted of 44 items, a mixture of vocabulary-in-context and extended passage comprehension item types. The grade 3 DRP test form (H-9) contained 42 items and all remaining grades (5, 7, and 10) consisted of 70 items on the DRP test.

Star Reading and DRP test score data were obtained on 273 students at grade 3, 424 students at grade 5, 353 students at grade 7, and 314 students at grade 10.

Item-level factor analysis of the combined Star and DRP response data indicated that the tests were essentially measuring the same construct at each of the four grades. Latent roots (Eigenvalues) from the factor analysis of the tetrachoric correlation matrices tended to verify the presence of an essentially unidimensional construct. In general, the Eigenvalue associated with the first factor was very large in relation to the eigenvalue associated with the second factor. Overall, these results confirmed the essential unidimensionality of the combined Star Reading and DRP data. Since DRP is an acknowledged measure of reading comprehension, the factor

analysis data support the claim that Star Reading likewise measures reading comprehension.

Subsequent to the factor analysis, the Star Reading item difficulty parameters were transformed to the DRP difficulty scale, so that scores on both tests could be expressed on a common scale. Star Reading scores on that scale were then calculated using the methods of Item Response Theory. Table 38 below shows the correlations between Star Reading and DRP reading comprehension scores overall and by grade.

Table 38: Correlations between Star Reading and DRP Test Scores, Overall and by Grade

Grade	Sample Size	Test Form		Number of Items		Correlation
		Star Calibration	DRP	Star	DRP	
3	273	321	H-9	44	42	0.84
5	424	511	H-7	44	70	0.80
7	353	623	H-6	44	70	0.76
10	314	701	H-2	44	70	0.86
Overall	1,364					0.89

In summary, using item factor analysis Yoes (1999) showed that Star Reading items measure the same underlying construct as the DRP: reading comprehension. The overall correlation of 0.89 between the DRP and Star Reading test scores corroborates that. Furthermore, correcting that correlation coefficient for the effects of less than perfect reliability yields a corrected correlation of 0.96. Thus, both at the item level and at the test score level, Star Reading was shown to measure essentially the same construct as the DRP.

Investigating Oral Reading Fluency and Developing the Estimated Oral Reading Fluency Scale

During the fall of 2007 and winter of 2008, 32 schools across the United States that were then using both Star Reading and DIBELS oral reading fluency (DORF) for interim assessments participated in a research study to evaluate the relationship of Star Reading scores to oral reading fluency. Below are highlights of the methodology and results of the study.

A single-group design provided data for both evaluation of concurrent validity and the linking of the two score scales. For the linking analysis, an equipercentile methodology was used. Analysis was done independently for each of grades 1–4. To evaluate the extent to which the linking accurately approximated student performance, 90 percent of the sample was used

to calibrate the linking model, and the remaining 10 percent were used for cross-validating the results. The 10 percent were chosen by a simple random function.

The 32 schools in the sample came from 9 states: Alabama, Arizona, California, Colorado, Delaware, Illinois, Michigan, Tennessee, and Texas. This represented a broad range of geographic areas, and resulted in a large number of students (N = 12,220). The distribution of students by grade was as follows:

- 1st grade: 2,001
- 2nd grade: 4,522
- 3rd grade: 3,859
- 4th grade: 1,838

The sample was composed of 61 percent of students of European ancestry; 21 percent of African ancestry; 11 percent of Hispanic ancestry; with the remaining 7 percent of Native American, Asian, or other ancestry.

Students were individually assessed using the DORF (DIBELS Oral Reading Fluency) benchmark passages. The students read the three benchmark passages under standardized conditions. The raw score for passages was computed as the number of words read correctly within the one-minute limit (WCPM, Words Correctly read Per Minute) for each passage. The final score for each student was the median WCPM across the benchmark passages, and was the score used for analysis. Each student also took a Star Reading assessment within two weeks of the DORF assessment.

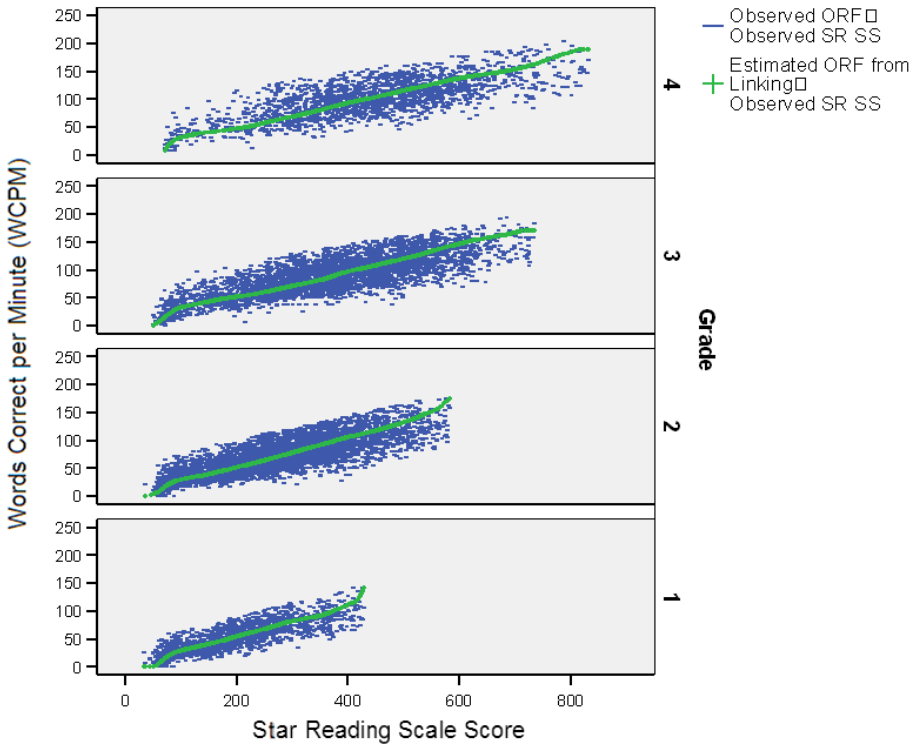
Descriptive statistics for each grade in the study on Star Reading Scaled Scores and DORF WCPM (words correctly read per minute) are found in Table 39.

Correlations between the Star Reading Scaled Score and DORF WCPM at all grades were significant ($p < 0.01$) and diminished consistently as grades increased. Figure 4 visualizes the scatterplot of observed DORF WCPM and SR Scaled Scores, with the equipercentile linking function overlaid. The equipercentile linking function appeared linear; however, deviations at the tails of the distribution for higher and lower performing students were observed. The root mean square errors of linking for grades 1–4 was found to be 14, 19, 22, and 25 WCPM, respectively.

Table 39: Descriptive Statistics and Correlations between Star Reading Scale Scores and DIBELS Oral Reading Fluency for the Calibration Sample

Grade	N	Star Reading Scale Score		DORF WCPM		Correlation
		Mean	SD	Mean	SD	
1	1,794	172.90	98.13	46.05	28.11	0.87
2	4,081	274.49	126.14	72.16	33.71	0.84
3	3,495	372.07	142.95	90.06	33.70	0.78
4	1,645	440.49	150.47	101.43	33.46	0.71

Figure 4: Scatterplot of Observed DORF WCPM and SR Scale Scores for Each Grade with the Grade Specific Linking Function Overlaid



Cross-Validation Study Results

The 10 percent of students randomly selected from the original sample were used to provide evidence of the extent to which the models based on the calibration samples were accurate. The cross-validation sample was intentionally kept out of the calibration of the linking estimation, and the results of the calibration sample linking function were then applied to the cross-validation sample.

Table 40 provides descriptive information on the cross-validation sample. Means and standard deviations for DORF WCPM and Star Reading Scaled Score for each grade were of a similar magnitude to the calibration sample. Table 41 provides results of the correlation between the observed DORF WCPM scores and the estimated WCPM from the equipercentile linking. All correlations were similar to results in the calibration sample. The average differences between the observed and estimated scores and their standard deviations are reported in Table 41 along with the results of one sample t-test evaluating the plausibility of the mean difference being significantly different from zero. At all grades the mean differences were not significantly different from zero, and standard deviations of the differences were very similar to the root mean square error of linking from the calibration study.

Table 40: Descriptive Statistics and Correlations between Star Reading Scale Scores and DIBELS Oral Reading Fluency for the Cross-Validation Sample

Grade	N	Star Reading Scale Score		DORF WCPM	
		Mean	SD	Mean	SD
1	205	179.31	100.79	45.61	26.75
2	438	270.04	121.67	71.18	33.02
3	362	357.95	141.28	86.26	33.44
4	190	454.04	143.26	102.37	32.74

Table 41: Correlation between Observed WCPM and Estimated WCPM Along with the Mean and Standard Deviation of the Differences between Them

Grade	N	Correlation	Mean Difference	SD Difference	t-test on Mean Difference
1	205	0.86	−1.62	15.14	$t(204) = -1.54$, $p = 0.13$
2	438	0.83	0.23	18.96	$t(437) = 0.25$, $p = 0.80$
3	362	0.78	−0.49	22.15	$t(361) = -0.43$, $p = 0.67$
4	190	0.74	−1.92	23.06	$t(189) = -1.15$, $p = 0.25$

Classification Accuracy of Star Reading

Accuracy for Predicting Proficiency on a State Reading Assessment

Star Reading test scores have been linked statistically to numerous state reading assessment scores. The linked values have been employed to use

Star Reading to predict student proficiency in reading on those state tests. One example of this is a linking study conducted using a multi-state sample of students' scores on the PARCC consortium assessment[2]. The table below presents classification accuracy statistics for grades 3 through 8.

Table 42: Classification diagnostics for predicting students' reading proficiency on the PARCC consortium assessment from earlier Star Reading scores

Measure	Grade					
	3	4	5	6	7	8
Overall classification accuracy	86%	87%	86%	86%	86%	83%
Sensitivity	64%	73%	73%	69%	73%	70%
Specificity	93%	93%	90%	91%	91%	89%
Observed proficiency rate (OPR)	26%	29%	27%	24%	28%	29%
Projected proficiency rate (PPR)	22%	26%	26%	23%	27%	28%
Proficiency status projection error	–5%	–3%	0%	–1%	–1%	–1%
Area under the ROC curve	0.91	0.93	0.91	0.92	0.92	0.90

As the table shows, classification accuracy ranged from 83 to 87%, depending on grade. Area Under the Curve (AUC) was at least 0.90 for all grades. Specificity was especially high, and the projected proficiency rates were very close to the observed proficiency rates at all grades.

Numerous other reports of linkages between Star Reading and state accountability tests have been conducted. Reports are available at http://research.renaissance.com/.

Accuracy for Identifying At-Risk Students

In many settings, Star Reading is used to identify students considered "at risk" for reading difficulties requiring intervention, for example long in advance of state accountability assessment that will be used to classify students at the end of the school year. This section summarizes two studies done to evaluate the validity of cut scores based on Star Reading as predictors of "at risk" status later in the school year. In such cases, correlation coefficients are of less interest than classification accuracy statistics, such as overall accuracy of classification, sensitivity and specificity, false positives and false negatives, positive and negative predictive power, receiver operating characteristic

2. Renaissance Learning (2016). Relating Star Reading™ and Star Math™ to the Colorado Measure of Academic Success (CMAS) (PARCC Assessments) performance.

(ROC) curves, and a summary statistic called AUC (Area Under the Curve).[3] Summaries of the methodology and results of the two studies are given below.

Brief Description of the Current Sample and Procedure

Initial Star Reading classification analyses were performed using state assessment data from Arkansas, Delaware, Illinois, Michigan, Mississippi, and Kansas. Collectively these states cover most regions of the country (Central, Southwest, Northeast, Midwest, and Southeast). Both the Classification Accuracy and Cross Validation study samples were drawn from an initial pool of 79,045 matched student records covering grades 2–11.

A secondary analysis using data from a single state assessment was then performed. The sample used for this analysis was 42,771 matched Star Reading and South Dakota Test of Education Progress records of students in grades 3–8.

An ROC analysis was used to compare the performance data on Star Reading to performance data on state achievement tests, with "at risk" identification as the criterion. The Star Reading Scaled Scores used for analysis originated from assessments 3–11 months before the state achievement tests were administered. Selection of cut scores was based on the graph of sensitivity and specificity versus the Scaled Score. For each grade, the Scaled Score chosen as the cut point was equal to the score where sensitivity and specificity intersected. The classification analyses, cut points and outcome measures are outlined in Table 43. Area Under the Curve (AUC) values were all greater than 0.80. Descriptive notes for other values represented in the table are provided in the table footnote.

3. For descriptions of ROC curves, AUC, and related classification accuracy statistics, refer to Pepe, Janes, Longton, Leisenring, & Newcomb (2004) and Zhou, Obuchowski & Obushcowski (2002.)

Table 43: Classification Accuracy in Predicting Proficiency on State Achievement Tests in Seven States[a]

Statistic[b]	Initial Analysis		Secondary Analysis	
	Value		Value	
False Positive Rate	21%		18%	
False Negative Rate	23%		22%	
Sensitivity	76%		78%	
Specificity	76%		82%	
Overall Classification Rate	76%		81%	
	Grade	AUC	Grade	AUC
AUC (ROC)	2	0.816		
	3	0.839	3	0.869
	4	0.850	4	0.882
	5	0.841	5	0.881
	6	0.833	6	0.883
	7	0.829	7	0.896
	8	0.843	8	0.879
	9	0.847		
	10	0.858		
	11	0.840		
	10	777		
	11	1,055		

a. Arkansas, Delaware, Illinois, Kansas, Michigan, Mississippi, and South Dakota.
b. The false positive rate is equal to the proportion of students incorrectly labeled "at-risk." The false negative rate is equal to the proportion of students incorrectly labeled not "at-risk." Likewise, sensitivity refers to the proportion of correct positive predictions while specificity refers to the proportion of negatives that are correctly identified (e.g., student will not meet a particular cut score).

Disaggregated Validity and Classification Data

In some cases, there is a need to verify that tests, such as Star Reading, as valid for different demographic groups. For that purpose, the data must be disaggregated, and separate analyses performed for each group. Table 44 shows the disaggregated classification accuracy data for ethnic subgroups and also the disaggregated validity data.

Table 44: Disaggregated Classification and Validity Data

Classification Accuracy in Predicting Proficiency on State Achievement Tests in 6 States (Arkansas, Delaware, Illinois, Kansas, Michigan, and Mississippi): by Race/Ethnicity										
	White, non-Hispanic (n = 17,567)		**Black, non-Hispanic (n = 8,962)**		**Hispanic (n = 1,382)**		**Asian/Pacific Islander (n = 231)**		**American Indian/Alaska Native (n = 111)**	
False Positive	31%		44%		36%		17%		12%	
False Negative Rate	38%		12%		12%		24%		41%	
Sensitivity	62%		88%		88%		76%		59%	
Specificity	87%		56%		64%		83%		88%	
Overall Classification Rate	81%		67%		73%		82%		78%	
AUC (ROC)	**Grade**	**AUC**	**Grade**	**AUC**	**Grade**	**AUC**	**Grade**	**AUC**	**Grade**	**AUC**
	2	n/a	2	0.500	2	n/a	2	n/a	2	n/a
	3	0.863	3	0.828	3	0.868	3	0.913	3	0.697
	4	0.862	4	0.823	4	0.837	4	0.869	4	0.888
	5	0.853	5	0.832	5	0.839	5	0.855	5	0.919
	6	0.849	6	0.806	6	0.825	6	0.859	6	0.846
	7	0.816	7	0.784	7	0.866	7	0.904	7	0.900
	8	0.850	8	0.827	8	0.812	8	0.961	8	1.000
	9	1.000	9	0.848	9	n/a	9	n/a	9	n/a
	10	0.875	10	0.831	10	0.833	10	n/a	10	n/a
	11	0.750	11	1.000	11	n/a	11	n/a	11	n/a

Evidence of Technical Accuracy for Informing Screening and Progress Monitoring Decisions

Many school districts use tiered models such as Response to Intervention (RTI) or Multi-Tiered Systems of Support (MTSS) to guide instructional decision making and improve outcomes for students. These models represent a more proactive, data-driven approach for better serving students as compared with prior decision-making practices, including processes to:

- Screen all students to understand where each is in the progression of learning in reading, math, or other disciplines
- Identify at-risk students for intervention at the earliest possible moment
- Intervene early for students who are struggling or otherwise at-risk of falling behind; and
- Monitor student progress in order to make decisions as to whether they are responding adequately to the instruction/intervention

Assessment data are central to both screening and progress monitoring, and Star Reading is widely used for both purposes. This chapter includes technical information about Star Reading's ability to accurately screen students according to risk and to help educators make progress monitoring decisions. Much of this information has been submitted to and reviewed by the Center on Response to Intervention https://rti4success.org/ and/or the National Center on Intensive Intervention https://intensiveintervention.org/, two technical assistance groups funded by the US Department of Education.

For several years running, Star Reading has enjoyed favorable technical reviews for its use in informing screening and progress monitoring decision by the CRTI and NCII, respectively. The most recent reviews by CRTI indicate that Star Reading has a "convincing" level of evidence (the highest rating awarded) in the core screening categories, including classification accuracy, reliability, and validity. CRTI also notes that the extent of the technical evidence is "Broad" (again, the highest rating awarded) and notes that not only is the overall evidence compelling, but there are disaggregated data as well that shows Star Reading works equally well among subgroups. The most recent reviews by NCII indicate that there is full "convincing" evidence of Star Reading's psychometric quality for progress monitoring purposes, including reliability, validity, reliability of the slope, and validity of the slope. Furthermore, they find fully "convincing" evidence that Star Reading is sufficiently sensitive to student growth, has adequate alternate forms, and provides data-based guidance to educators on end-of-year benchmarks and when an intervention should be changed, among other categories. Readers may find additional information on Star Reading on those sites and should note that the reviews are updated on a regular basis, as their review standards are adjusted and new technical evidence for Star Reading and other assessments are evaluated.

Screening

According to the Center on Response to Intervention, "Screening is conducted to identify or predict students who may be at risk for poor learning outcomes. Universal screening assessments are typically brief, conducted with all students at a grade level, and followed by additional testing or short-term progress monitoring to corroborate students' risk status."[4]

Most commonly, screening is conducted with all students at the beginning of the year and then another two to four times throughout the school year. Star

4. https://rti4success.org/essential-components-rti/universal-screening

Reading is widely used for this purpose. In this section, the technical evidence supporting its use to inform screening decisions is summarized.

Organizations of RTI/MTSS experts such as the Center on Response to Intervention and the RTI Action Network[5] are generally consistent in how measurement tools should be evaluated for their appropriateness as screeners. Key categories include the following:

1. **Validity and reliability.** Data on Star Reading's reliability were presented in the "Reliability and Measurement Precision" chapter of this manual. A wide array of validity evidence has been presented in this chapter, above; detailed tables of correlational data can be found in "Appendix B: Detailed Evidence of Star Reading Validity".
2. **Practicality and efficiency.** Screening measures should not require much teacher or student time. Because most students can complete a Star Reading test in 15–20 minutes or less, and because it is group administered and scored automatically, Star Reading is an exceptionally efficient general outcomes measure for reading.
3. **Classification accuracy metrics** including sensitivity, specificity, and overall predictive accuracy. These are arguably the most important indicators, addressing the main purpose of screening: When a brief screening tool indicates a student either is or is not at risk of later reading difficulties, how often is it accurate, and what types of errors are made?

It is common to use high stakes indicators such as state summative assessments as criterion measures for classification accuracy evaluation. Star Reading is linked to virtually every state summative assessment in the US as well as the ACT and SAT college entrance exams. The statistical linking of the Star Reading scale with these other measures' scales, combined with Star Reading growth norms (discussed in the Norming chapter of this manual) empowers Star Reading reports and data extracts to make predictions throughout the school year about future student performance. These predictions inform educator screening decisions in schools using an RTI/MTSS framework. (Educators are also free to use norm-referenced scores such as Percentile Ranks to inform screening decisions.)

Star Reading's classification accuracy results from several recent predictive studies are summarized in Table 45. Each study evaluated the extent to which Star Reading accurately predicted whether a student achieved a specific performance level on another reading or English Language Arts measure. The specific performance level (cut point) varies by assessment and grade. Cut

5. http://www.rtinetwork.org/learn/research/universal-screening-within-a-rti-model

points are set by assessment developers and sponsors, which in the case of state summative exams usually means the state department of education and/or state board of education. State assessments generally have between three and five performance levels, and the cut point used in these analyses refers to the level the state has determined indicates meeting grade level reading or English Language Arts standards. For instance, the cut point on California's CAASPP is Level 3, also known as "Standard Met." On Louisiana's LEAP 2025 the cut point is at the "Mastery" level. In the case of ACT and SAT, the cut point established by the developers (ACT and College Board, respectively) indicates an estimated level of readiness for success in college.

Table 45: Summary of classification accuracy metrics from recent studies linking Star Reading with summative reading and English Language Arts measures

Assessment	Grade/s Covered	Date study completed	Study sample size	Average result across all grades			
				Overall classification accuracy	Sensitivity	Specificity	Area under ROC curve
ACT English (college readiness)	11	4/22/2016	14,248	80%	76%	82%	0.87
ACT Reading (college readiness)	11	4/22/2016	14,228	83%	62%	90%	0.86
ACT Aspire	3–10	6/1/2017	44,877	84%	81%	84%	0.92
California Assessment of Student Performance and Progress (CAASPP) (Smarter Balanced)	3–8	10/30/2015	51,835	84%	86%	82%	0.92
Florida Standards Assessments (FSA)	3–8	6/30/2015	41,178	84%	84%	83%	0.92
Georgia Milestones	3–8	7/1/2017	44,436	87%	79%	90%	0.94
Illinois Partnership for Assessment of Readiness for College and Careers (PARCC) Assessments	3–10	7/13/2016	27,415	86%	70%	91%	0.91
Louisiana Educational Assessment Program (LEAP 2025)	3–8	12/1/2017	33,815	84%	90%	69%	0.90
Maine Educational Assessment (MEA)	3–8	7/1/2017	945	83%	78%	86%	0.93
Mississippi Academic Assessment Program (MAAP)	3–8	2/1/2017	13,590	84%	80%	87%	0.92

Table 45: Summary of classification accuracy metrics from recent studies linking Star Reading with summative reading and English Language Arts measures

Assessment	Grade/s Covered	Date study completed	Study sample size	Average result across all grades			
				Overall classification accuracy	Sensitivity	Specificity	Area under ROC curve
Missouri Assessment Program (MAP) Grade-Level Tests	3–8	3/14/2017	30,626	85%	83%	87%	0.96
North Carolina READY End-of-Grade (EOG)	3–8	2/16/2015	396,075	81%	83%	78%	0.89
Ohio State Tests	3–8	12/20/2016	27,487	85%	83%	87%	0.93
Pennsylvania's System of School Assessment (PSSA)	3–8	12/19/2016	7,383	85%	91%	72%	0.92
SAT (college entrance)	11						
South Carolina College-and Career-Ready Assessments (SC READY)	3–8	12/5/2016	10,011	86%	85%	86%	0.94
State of Texas Assessments of Academic Readiness (STAAR)	3–8	7/1/2017	3,915	83%	71%	88%	0.90
Wisconsin Forward Exam	3–8	12/22/2016	39,605	88%	73%	93%	0.94

Notes:

- Some tests, such as the Smarter Balanced (indicated above for California) and PARCC (indicated above for Illinois) are used in multiple states, so those results may apply to other states not listed here.
- Overall classification accuracy refers to the percentage of correct classifications.
- Sensitivity refers to the rate at which Star Reading identifies students as being at-risk who demonstrate a poor learning outcome at a later point in time. Sensitivity can be thought of as the true positive rate. Screening tools with high sensitivity help ensure that students who truly need intervention will be identified to receive it.
- Specificity refers to the rate at which Star Reading identifies students as being not at-risk who perform satisfactorily at a later point in time. Specificity can be thought of as a true negative rate. Screening tools with high specificity help ensure that scarce resources are not invested in students who do not require extra assistance.
- Area under the ROC (Receiver Operating Characteristic) curve is a powerful indicator of overall accuracy. The ROC curve a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for the full range of possible screener (Star Reading) cut points. The area under ROC Curve (AUC) is an overall indication of the diagnostic accuracy of the curve. AUC values range between 0 and 1 with 0.5 indicating a chance level of accuracy. The Center for Response to Intervention considers results at or above 0.85 to be an indication of convincing evidence of classification accuracy.[6]

6. https://rti4success.org/resources/tools-charts/screening-tools-chart/screening-tools-chart-rating-system

Note that many states tend to not use the same assessment system for more than a few consecutive years, and Renaissance endeavors to keep the Star Reading classification reporting as up to date as possible. Those interested in reviewing the full technical reports for these or other state assessments are encouraged to visit http://research.renaissance.com/advancedsearch.asp and search by state name for the Star Reading linking reports (e.g., "Wisconsin linking").

Progress Monitoring

According to the National Center on Intensive Intervention, "progress monitoring is used to assess a student's performance, to quantify his or her rate of improvement or responsiveness to intervention, to adjust the student's instructional program to make it more effective and suited to the student's needs, and to evaluate the effectiveness of the intervention."[7]

In an RTI/MTSS context, progress monitoring involves frequent assessment—usually occurring once every 1–4 weeks—and often involves only those students who are receiving additional instruction after being identified as at-risk via the screening process. Ultimately, educators use progress monitoring data to determine whether a student is responding adequately to the instruction, or whether adjustments need to be made to the instructional intensity or methods. The idea is to get to a decision quickly, with as little testing as possible, so that valuable time is not wasted on ineffective approaches. Educators make these decisions by comparing their performance against a goal set by the educator. Goals should be "reasonable yet ambitious"[8] as recommended by Shapiro (2008), and Star Reading offers educators a variety of guidance to set normative or criterion-referenced goals that meet these criteria.

The RTI Action Network, National Center on Intensive Intervention, and other organizations offering technical assistance to schools implementing RTI/MTSS models are generally consistent in encouraging educators to select assessments for progress monitoring that have certain characteristics.

7. https://intensiveintervention.org/ncii-glossary-terms#ProgresMonitoring
8. Shapiro, E. S. (2008). Best practices in setting progress-monitoring goals for academic skill improvement. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 141-157). Bethesda, MD: National Association of School Psychologists.

A summary of those characteristics and relevant information about Star Reading is provided below.

1. **Evidence of psychometric quality.**

   a. **Reliability and validity.** Summaries of the available evidence supporting Star Reading's reliability and validity are presented in the chapter on "Reliability and Measurement Precision" and throughout this Validity chapter.

   b. **Reliability of the slope.** Because progress monitoring decisions often involve the student's rate of progress over multiple test administrations, the characteristics of the student's slope of improvement, or trend line, are also important. A study was conducted in 2017 by Renaissance Learning to evaluate reliability of slope for at-risk students who were being progress monitored during the 2016–17 school year. Specifically, the sample included 218,689 students who began the year below the 30th Percentile Rank in Star Reading and were assessed 10 or more times during the school year, with a minimum of 140 days between first and last test.

   Every student's Star Reading test records were sorted in chronological order. Each test record was coded as either an odd- or even-numbered test. Slopes were estimated for each student's odd-number tests and also for the even-numbered tests using ordinary least squares regression. Then, the odd and even slopes were correlated. The table below summarizes the Pearson correlation coefficients by grade, indicating a consistently strong association between even and odd numbered test slopes.

Table 46: Star Reading Reliability of the Slope Coefficients by grade, 1–12

Grade	n	Coefficient
1	14,179	0.76
2	43,978	0.93
3	52,670	0.94
4	37,862	0.93
5	31,326	0.93
6	16,990	0.94
7	9,683	0.94
8	7,786	0.94
9	2,483	0.94
10	1,549	0.94
11	799	0.94
12	384	0.95

2. **Produce a sufficient number of forms.** Because Star Reading is computer adaptive and its item bank comprises more than six thousand items, there are at a minimum, several hundred alternate forms for a student at a given ability level. This should be more than sufficient for even the most aggressive progress monitoring testing schedule.

   A variety of grade-specific evidence is available to demonstrate the extent to which Star Reading can reliably produce consistent scores across repeated administrations of the same or similar tests to the same individual or group. These include:

   a. Generic reliability, defined as the proportion of test score variance that is attributable to true variation in the trait or construct the test measures.

   b. Alternate forms reliability, defined as the correlation between test scores on repeated administrations to the same examinees.

   Grade-level results are summarized in the "Reliability and Measurement Precision" chapter.

3. **Practicality and efficiency.** As mentioned above, most students complete Star Reading in 15–20 minutes. It is auto-scored and can be group administered, requiring very little educator involvement, making it an efficient progress monitoring solution.

4. **Specify criterion for adequate growth and benchmarks for end-of-year performance levels.** Goal-setting decisions are handled by local educators, who know their students best and are familiar with the efficacy and intensity of the instructional supports that will be offered. That said, publishers of assessments used for progress monitoring are expected to provide empirically based guidance to educators on setting goals.

   Star Reading provides guidance to inform goal setting using a number of different metrics, including the following:

   a. **Student Growth Percentile.** SGP describes a student's velocity (slope) relative to a national sample of academic peers—those students in the same grade with a similar score history. SGPs work like Percentile Ranks (1–99 scale) but once an SGP goal has been set, it is converted to a Scaled Score goal at the end date specified by the teacher. An SGP-defined goal can be converted into an average weekly increase in a Scaled Score metric, if educators prefer to use that. Many teachers select either SGP 50 (indicating typical or expected growth) as minimum acceptable growth, or something indicating accelerated growth, such as 65 or 75. A helpful feature of SGP is that it can be used as a "reality check" for any goal, whether it be in an SGP metric or something else (e.g., Scaled Score, Percentile Rank). SGP estimates the likelihood that the student will achieve a level of growth or later performance. For example, a goal associated with an SGP of 75

indicates that only about 25 percent of the student's academic peers would be expected to achieve that level of growth.

b. **State test proficiency.** As described in the Screening section, the fact that Star Reading is linked to virtually every state assessment enables educators to select values on the Star scale that are approximately equivalent to states' defined proficiency level cut points for each grade.

c. **Percentile Rank and Scaled Score.** Educators may also enter custom goals using Percentile Rank or Scaled Score metrics.

Additional research on Star Reading as a progress monitoring tool

A 2017 study by Cormier & Bulut (manuscript under review)[9] evaluated Star Reading as a progress monitoring tool, concluding:

- Although relatively little research exists on using computer adaptive measures for progress monitoring as opposed to curriculum based measurement probes, the study concluded it was possible to use Star Reading for progress monitoring purposes.
- Sufficiently reliable progress monitoring slopes could be generated in as few as five Star Reading administrations.
- The duration of Star Reading progress monitoring (i.e., over how many weeks) should be conducted is a function of the amount of typical growth by grade in relation to measurement error. For earlier grades (when student rates of growth are greatest), that amount of time could be as little as six weeks. For middle grades, 20 weeks should be sufficient.
- These two findings challenge popular rules of thumb about progress monitoring frequency and duration (most of which are derived from CBM probe studies), which often involve weekly testing over periods of time that are selected due to popular convention rather than empirical evidence.
- Using Theil-Sen regression procedures to estimate slope as opposed to OLS could reduce the influence of outlier scores, and thus provide a more accurate picture of student growth.

Summary of Star Reading Validity Evidence

The validity data presented in this technical manual includes evidence of Star Reading's concurrent, predictive, and construct validity, as well as

9. Cormier, D. & Bulut, O. (Manuscript under review.) Using computer-adaptive tests to monitor student progress: optimal slope estimates, administration frequency, and data points.

classification accuracy statistics, and strong measures of association with non-traditional reading measures such as oral reading fluency. Exploratory and confirmatory factor analyses provided evidence that Star Reading measures a unidimensional construct, consistent with the assumption underlying its use of the Rasch 1-parameter logistic item response model. The Meta-Analysis section showed the average uncorrected correlation between Star Reading and all other reading tests to be 0.79. (Many meta-analyses adjust the correlations for range restriction and attenuation to less than perfect reliability; had we done that here, the average correlation would have exceeded 0.85.) Correlations with specific measures of reading ability were often higher than this average. For example, Yoes (1999) found within-grade correlations with DRP averaging 0.81. When these data were combined across grades, the correlation was 0.89. The latter correlation may be interpreted as an estimate of the overall construct validity of Star Reading as a measure of reading comprehension. Yoes also reported that results of item factor analysis of DRP and Star Reading items yielded a single common unidimensional factor. This provides strong support for the claim that Star Reading is a measure of reading comprehension.

International data from the UK show even stronger correlations between Star Reading and widely used reading measures there: overall correlations of 0.91 with the Suffolk Reading Scale, median within-form correlations of 0.84, and a correlation of 0.85 with teacher assessments of student reading.

Finally, the data showing the relationship between the current, standards-based Star Reading Enterprise test and scores on specific state accountability tests and on the SBAC and PARCC Common Core consortium tests show that the correlations with these summative measures are consistent with the meta-analysis findings.

References

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

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

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

Norming

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

Background

National norms for Star Reading version 1 were first collected in 1996. Substantial changes introduced in Star Reading version 2 necessitated the development of new norms in 1999. Those norms were used until new norms were developed in 2008. Since 2008, Star Reading norms have been updated twice (2014 and 2017). The 2017 norms went live in Star in the 2017–2018 school year. This chapter describes the development of the 2017 norms.

From 1996 through mid-2011, Star Reading was primarily a measure of reading comprehension comprising short vocabulary-in-context items and longer passage comprehension items. The current version of Star Reading, introduced in June 2011, is a standards-based assessment that measures a wide variety of skills and instructional standards, as well as reading comprehension. To develop the current version of Star Reading, scale scores were equated to the scale used in earlier versions of Star Reading. The equating analyses demonstrated that, despite its distinctive content, the latent attribute underlying the current version is the same one underlying previous versions of Star Reading. It measures the same broad construct, and reports student performance on the same score scale. As part of the 2014 norming process, scores from the older version of Star Reading were equated to the current version and the 2014 norms were applied both to the current and original versions of Star Reading.

The 2017 Star Reading Norms

Prior to development of the 2017 Star Reading norms, a new reporting scale was developed, called the Unified scale. The Unified scale is a linear transformation of Star Reading's Rasch ability scale to a new integer scale that is also used in Star Early Literacy. The Star Unified scale makes it

possible to report performance on both of those Star assessments on the same scale.

The original Star Reading scale, now referred to as the "Enterprise" score scale, was based on a nonlinear transformation of Rasch scores. Both the Enterprise and the Unified scale scores will be available to Star test users during the planned transition to the Unified scale as the default reporting scale.

New U.S. norms for Star Reading assessments (Early Literacy and Reading) were introduced at the start of the 2017–18 school year. Separate early fall and late spring norms were developed for grades Kindergarten through 12. In previous Star Reading norming analyses, the reference populations for grades Kindergarten through 3 consisted only of students taking Star Reading; students who only took Star Early Literacy were excluded from the Star Reading norms, and vice versa. Consequently, previous Star Reading norms for this grade range were not completely representative of the full range of literacy development in those grades. To address this, the concept of "Star Early Learning" was introduced. That concept acknowledges the overlap of literacy development content between the Star Reading and Early Literacy assessments, and encompasses in the normative reference group all students in each of grades K–3 who have taken either the Reading assessment, the Early Literacy assessment, or both.

The norms introduced in 2017 are based on test scores of K–3 students taking either the Reading assessment, or the Early Literacy one, or both. These norms are based on use of the Unified scale, which allowed performance on both Star Early Literacy and Star Reading to be measured on the same scale.

Students participating in the norming study took assessments between August 15, 2014 and June 30, 2015. Students took the Star Reading or Early Literacy tests under normal test administration conditions. No specific norming test was developed and no deviations were made from the usual test administration. Thus, students in the norming sample took Star Reading tests as they are administered in everyday use.

Sample Characteristics

During the norming period, a total of 5,814,221 US students in grades K–12 took current Star Reading and/or Early Literacy tests administered using Renaissance Place servers hosted by Renaissance Learning. The first step in sampling was to select representative Fall and Spring student samples:

Students who had tested in the fall, in the spring, or in both the fall and spring of the 2014–2015 school year. From the fall and the spring samples, stratified subsamples were randomly drawn based on student grade and ability decile. The grade and decile sampling was necessary to ensure adequate and similar numbers of students in each grade, and each decile within grade. Because these norming data were convenience samples drawn from the Star Reading customer base, steps were taken to ensure the resulting norms were nationally representative of grades K–12 US student populations with regard to certain important characteristics. A post-stratification procedure was used to adjust the sample's proportions to the approximate national proportions on three key variables: geographic region, district socio-economic status, and district/school size. These three variables were chosen because they had previously been used in Star Reading norming studies to draw nationally representative samples, are known to be related to test scores, and were readily available for the schools in the Renaissance hosted database.

The final norming sample size, after selecting only students with test scores in either the fall or the spring or both fall and spring in the norming year, and further sampling by grade and ability decile was 3,699,263 students in grades K–12. There were 2,786,680 students in the fall norming sample and 1,855,730 students in the spring norming sample; 943,147 students were included in both norming samples. These students came from 18,113 schools across 50 states and the District of Columbia.

Table 47 and Table 48 provide a breakdown of the number of students participating per grade in the fall and in the spring, respectively.

Table 47: Numbers of Students per Grade in the Fall Norms Sample

Grade	N	Grade	N	Grade	N	Grade	N
K	212,035	4	447,754	8	94,691	12	18,092
1	340,079	5	364,271	9	25,063	Total	2,786,680
2	456,566	6	219,348	10	35,198		
3	419,912	7	128,011	11	25,660		

Table 48: Numbers of Students per Grade in the Spring Norms Sample

Grade	N	Grade	N	Grade	N	Grade	N
K	196,720	4	308,040	8	43,980	12	4,230
1	237,360	5	244,750	9	25,240	Total	1,855,730
2	264,790	6	125,070	10	22,720		
3	299,620	7	73,830	11	9,380		

National estimates of US student population characteristics were obtained from two entities: the National Center for Educational Statistics (NCES) and Market Data Retrieval (MDR).

- National population estimates of students' demographics (ethnicity and gender) in grades K–12 were obtained from NCES; these estimates were from the 2013–14 school year, the most recent data available. National estimates of race/ethnicity were computed using the NCES data based on single race/ethnicity and also a multiple race category. The NCES data reflect the most recent census data from the US census bureau.
- National estimates of school-related characteristics were obtained from May 2016 Market Data Retrieval (MDR) information. The MDR database contains the most recent data on schools, some of which may not be reflected in the NCES data.

Table 49 on page 94 shows national percentages of children in grades K–12 by region, school/district enrollment, district socio-economic status, and location, along with the corresponding percentages in the fall and in the spring norming samples. MDR estimates of geographic region were based on the four broad areas identified by the National Educational Association as Northeastern, Midwestern, Southeastern, and Western regions. The specific states in each region are shown below.

Geographic Region

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

Northeast:

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

Southeast:

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

Midwest:

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

West:

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

School Size

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

Socioeconomic status as indexed by the percent of school students with free and reduced lunch

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

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

No students were sampled from the schools that did not report the percent of school students with free and reduced lunch.

The norming sample also included private schools, Catholic schools, students with disabilities, and English Language Learners as described below.

Table 49: Sample Characteristics Along with National Population Estimates and Sample Estimates

		National Estimates	Fall Norming Sample	Spring Norming Sample
Region	Midwest	20.9%	21.3%	19.3%
	Northeast	19.2%	10.9%	13.5%
	Southeast	24.6%	34.0%	29.7%
	West	35.3%	33.9%	37.6%
School Enrollment	< 200	4.0%	3.4%	3.6%
	200–499	26.8%	38.4%	38.6%
	≥ 500	69.1%	58.1%	57.8%
District Socioeconomic Status	Low	19.5%	24.8%	25.9%
	Below Median	24.3%	29.3%	27.7%
	Above Median	25.2%	23.0%	22.9%
	High	31.1%	22.9%	23.5%
Location	Rural	14.1%	20.8%	19.8%
	Suburban	42.3%	37.4%	38.0%
	Town	11.7%	16.5%	17.1%
	Urban	31.9%	25.4%	25.1%

Table 50 provides information on the demographic characteristics of students in the sample and national percentages provided by NCES. No weighting was done on the basis of these demographic variables; they are provided to help describe the sample of students and the schools they attended. Because Star assessment users do not universally enter individual student demographic information such as gender and ethnicity/race, some students were missing demographic data; the sample summaries in Table 50 are based on only those students for whom gender and ethnicity information were available. In addition to the student demographics shown, an estimated 7.4% of the students in the norming sample were gifted and talented (G&T)[1] as approximated by the 2011–2012 school data collected by the Office of Civil Rights (OCR). OCR is a subsidiary of the US Department of Education. School type was defined to be either public (including charter schools) or non-public (private, Catholic).

1. This estimate is based on data from the previous version of Star Reading norms. Given the similarity of the user pools for those and the 2017 norms, the current percentage is expected to be approximately the same.

Table 50: Student Demographics and School Information: National Estimates and Sample Percentages

			National Estimate	Fall Norming Sample	Spring Norming Sample
Gender	Public	Female	48.6%	49.9%	49.5%
		Male	51.4%	50.1%	50.5%
	Non-Public	Female	–	50.7%	50.6%
		Male	–	49.3%	49.4%
Race/Ethnicity	Public	American Indian	1.0%	1.6%	1.6%
		Asian	5.3%	5.2%	5.4%
		Black	15.5%	18.0%	17.9%
		Hispanic	25.4%	20.6%	22.7%
		White	49.6%	54.7%	52.4%
		Multiple Race[a]	3.2%	–	–
	Non-Public	American Indian	0.5%	2.6%	3.9%
		Asian	6.6%	3.7%	3.5%
		Black	9.1%	15.6%	15.5%
		Hispanic	10.7%	19.2%	27.8%
		White	69.2%	58.9%	49.3%
		Multiple Race[a]	3.9%	–	–

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

Test Administration

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

Data Analysis

Student test records were compiled from the complete database of Star Reading and Early Literacy Renaissance Place users. Data were from a single school year from August 2014 to June 2015. Students' unified scale Rasch scores on their first Star Reading or Early Literacy test taken during the first and second months of the school year based on grade placement were used to compute norms for the fall; students' Rasch scores on the last

Star Reading or Early Literacy test taken during the 8th and 9th months of the school year were used to compute norms for the spring. Interpolation was used to estimate norms for times of the year between the first month in the fall and the last month in the spring. The norms were based on the distribution of Rasch scores for each grade.

As noted above, a post-stratification procedure was used to approximate the national proportions on key characteristics. Post stratification weights from the regional, district socio-economic status, and school size strata were computed and applied to each student's unified Rasch ability estimate. Norms were developed based on the weighted Rasch ability estimates and then transformed to unified as well as Enterprise Star Reading scaled scores.[2] Table 51 provides descriptive statistics for each grade with respect to the normative sample performance, in the Unified scaled score units. Table 52 provides descriptive statistics for each grade with respect to the normative sample performance, in the Enterprise scaled score units.

Table 51: Descriptive Statistics for Weighted Scaled Scores by Grade for the Norming Sample in the Unified Scale

	Fall Unified Scaled Scores				Spring Unified Scaled Scores			
Grade	N	Mean	Standard Deviation	Median	N	Mean	Standard Deviation	Median
K	212035	702	62	703	196720	796	65	793
1	340079	776	72	767	237360	857	69	856
2	456566	887	70	888	264790	939	66	942
3	419912	952	67	956	299620	987	64	990
4	447754	994	64	999	308040	1021	65	1023
5	364271	1031	65	1036	244750	1055	67	1058
6	219348	1063	67	1067	125070	1085	70	1089
7	128011	1087	71	1090	73830	1104	74	1108
8	94691	1109	73	1114	43980	1126	77	1130
9	25063	1128	78	1131	25240	1138	76	1143
10	35198	1138	75	1143	22720	1143	77	1150
11	25660	1143	75	1150	9380	1150	75	1157
12	18092	1153	76	1161	4230	1158	76	1165

2. As part of the development of the Star Early Learning Unified scale, Star Early Literacy Rasch scores were equated to the Star Reading Rasch scale. This resulted in a downward extension of the latter scale that encompasses the full range of both Star Early Literacy and Reading performance. This extended Rasch scale was employed to put all students' scores on the same scale for purposes of norms development.

Table 52: Descriptive Statistics for Weighted Scaled Scores by Grade for the Norming Sample in the Enterprise Scale

	Fall Enterprise Scaled Scores				Spring Enterprise Scaled Scores			
Grade	N	Mean	Standard Deviation	Median	N	Mean	Standard Deviation	Median
K	212035	57	26	57	196720	85	52	80
1	340079	77	43	75	237360	162	108	139
2	456566	219	135	219	264790	324	144	317
3	419912	353	152	362	299620	439	168	435
4	447754	455	171	465	308040	525	213	522
5	364271	558	233	570	244750	637	259	640
6	219348	672	284	684	125070	786	346	795
7	128011	795	351	811	73830	889	374	895
8	94691	906	370	921	43980	984	359	994
9	25063	999	359	1026	25240	1092	338	1116
10	35198	1090	336	1124	22720	1132	331	1167
11	25660	1129	326	1172	9380	1172	311	1204
12	18092	1186	308	1224	4230	1216	286	1244

Growth Norms

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

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

The growth model used by Star Assessments is Student Growth Percentile (Betebenner, 2009). SGPs were developed by Dr. Damian Betebenner, originally

in partnership with several state departments of education.[3] It should be noted that the initial development of SGP involved annual state summative tests with reasonably constrained testing periods within each state. Because Star tests may be taken at multiple times throughout the year, a number of adaptations to the original model were made. For more information about Star Reading SGPs, please refer to this overview: http://doc.renlearn.com/KMNet/R00571375CF86BBF.pdf

SGPs are norm-referenced estimates that compare a student's growth to that of his or her academic peers nationwide. Academic peers are defined as those students in the same grade with a similar score history. SGPs are generated via a process that uses quantile regression to provide a measure of how much a student changed from one Star testing window to the next relative to other students with similar score histories. SGPs range from 1–99 and are interpreted similarly to Percentile Ranks, with 50 indicating typical or expected growth. For instance, an SGP score of 37 means that a student grew as much or more than 37 percent of her academic peers.

The Star Reading SGP package also produces a range of future growth estimates. Those are mostly hidden from users but are presented in goal setting and related applications to help users understand what typical or expected growth looks like for a given student. They are particularly useful for setting future goals and understanding the likelihood of reaching future benchmarks, such as likely achievement of proficient on an upcoming state summative assessment.

At present, the Star Reading SGP growth norms are based on a sample of 9.8 million student records across grades 1–12, with grade-specific samples ranging from about 100,000 to 1.3 million.

Reference

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

3. Core SGP documentation and source code are publicly available at https://cran.r-project.org/web/packages/SGP/index.html.

Score Definitions

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

Types of Test Scores

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

- *Criterion-referenced scores* describe a student's performance relative to a specific content domain or to a standard. Such scores may be expressed either on a continuous score scale or as a classification. An example of a criterion-referenced score on a continuous scale is a percent-correct score, which expresses what proportion of test questions the student can answer correctly in the content domain. An example of a criterion-referenced classification is a proficiency category on a standards-based assessment: the student may be said to be "proficient" or not, depending on whether the student's score equals, exceeds, or falls below a specific criterion (the "standard") used to define "proficiency" on the standards-based test. The criterion-referenced score reported by Star Reading is the Instructional Reading Level, which compares a student's test performance to the 1995 updated vocabulary lists that are based on the EDL's Core Vocabulary list. The Instructional Reading Level is the highest grade level at which the student is estimated to comprehend 80 percent of the text written at that level.
- *Norm-referenced scores* compare a student's test results to the results of other students who have taken the same test. In this case, scores provide a relative measure of student achievement compared to the performance of a group of students at a given time. Percentile Ranks and Grade Equivalents are the two primary norm-referenced scores available in Star Reading software. Both of these scores are based on a comparison of a student's test results to the data collected during the 2017 national norming program.

Scaled Score (SS)

Star Reading 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, and in order to provide a basis for deriving

the norm-referenced scores, it is necessary to convert all the results of Star Reading tests to scores on a common scale. Star Reading 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 Scaled Scores. Two different score scales are now available in Star assessments: the original scale scores, which are referred to as "Enterprise" scale scores; and a new score, expressed on the "Unified" score scale, which was introduced with the 2017–2018 school year.

Enterprise Scale Scores

For Star Reading, the "Enterprise" scale scores are the same scores that have been reported continuously since Star Reading Version 1 was introduced, in 1996. Because Version 1 was not based on item response theory, its scores were expressed on an ad hoc vertical (developmental) scale related to the student's reading grade level; scale scores ranged from 50 to 1350. The use of item response theory was introduced into Star Reading Version 2. Beginning with that version, Star software calculated students' scores on the Rasch IRT ability scale. To maintain continuity with the non-IRT score scale used in Version 1, the Rasch ability scores were converted to scores on the original scale by means of an equipercentile equating transformation. At that time, the range of reported Enterprise scale scores was extended to 0 to 1400.

Unified Scale Scores

Many users of Star Reading use Star Early Literacy to assess their students until they are ready to take Star Reading itself. Until recently, Star Reading and Star Early Literacy used different score scales, making it difficult to monitor growth as students transitioned from one assessment to the other. To ameliorate that disparity in the two tests' score scales, Renaissance developed a single score scale that applies to both assessments: the Unified score scale. That development began with equating the two tests' underlying Rasch ability scales; the result was the "unified Rasch scale", which is a downward extension of the Rasch scale used in all Star Reading versions since the introduction of version 2. The next step was to develop an integer scale based on the unified Rasch scale, with scale scores anchored to important points on the original Enterprise score scales of both tests. The end result was a reported score scale that extends from 200 to 1400: Star Early Literacy Unified scale scores range from 200 to 1100; Star Reading Unified scale scores range from 600 to 1400. An added benefit of the Unified scale is an improvement in certain properties of the scale scores: Scores on

both tests are much less variable from grade to grade; measurement error is likewise less variable; and Unified score reliability is slightly higher than that of the Enterprise scores.

Grade Equivalent (GE)

A Grade Equivalent (GE) indicates the grade placement of students for whom a particular score is typical. If a student receives a GE of 10.7, this means that the student scored as well on Star Reading as did the typical student in the seventh month of grade 10. It does not necessarily mean that the student can read independently at a tenth-grade level, only that he or she obtained a Scaled Score as high as the average tenth-grade, seventh-month student in the norms group.

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

Star Reading Grade Equivalents range from 0.0–12.9+. The scale divides the academic year into 10 monthly increments, and is expressed as a decimal with the unit denoting the grade level and the individual "months" in tenths. Table 53 indicates how the GE scale corresponds to the various calendar months. For example, if a student obtained a GE of 4.6 on a Star Reading assessment, this would suggest that the student was performing similarly to the average student in the fourth grade at the sixth month (March) of the academic year. Because Star Reading norms are based on fall and spring score data only, monthly GE scores are derived through interpolation by fitting a curve to the grade-by-grade medians. Table 31 on page 66 contains the Scaled Score to GE conversions.

Table 53: Incremental Grade Placements per Month

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

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

The Grade Equivalent scale is not an equal-interval scale. For example, an increase of 50 Scaled Score points might represent only two or three

months of GE change at the lower grades, but over a year of GE change in the high school grades. This is because student growth in reading (and other academic areas) is not linear; it occurs much more rapidly in the lower grades and slows greatly after the middle years. Consideration of this should be made when averaging GE scores, especially if it is done across two or more grades.

Comparing the Star Reading Test with Conventional Tests

Because the Star Reading test adapts to the reading level of the student being tested, Star Reading GE scores are more consistently accurate across the achievement spectrum than those provided by conventional test instruments. Grade Equivalent scores obtained using conventional (non-adaptive) test instruments are less accurate when a student's grade placement and GE score differ markedly. It is not uncommon for a fourth-grade student to obtain a GE score of 8.9 when using a conventional test instrument. However, this does not necessarily mean that the student is performing at a level typical of an end-of-year eighth-grader; more likely, it means that the student answered all, or nearly all, of the items correctly and thus performed beyond the range of the fourth-grade test.

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

Estimated Oral Reading Fluency (Est. ORF)

Estimated Oral Reading Fluency (Est. ORF) is an estimate of a student's ability to read words quickly and accurately in order to comprehend text efficiently. Students with oral reading fluency demonstrate accurate decoding, automatic word recognition, and appropriate use of the rhythmic aspects of language (e.g., intonation, phrasing, pitch, and emphasis).

Est. ORF is reported as an estimated number of words a student can read correctly within a one-minute time span on grade-level-appropriate text. Grade-level text was defined to be connected text in a comprehensible passage form that has a readability level within the range of the first half of

the school year. For instance, an Est. ORF score of 60 for a second-grade student would be interpreted as meaning the student is expected to read 60 words correctly within one minute on a passage with a reading grade level between 2.0 and 2.5. Therefore, when this estimate is compared to an observed score on a specific passage which has a fixed level of readability, there might be noticeable differences as the Est. ORF provides an estimate across a range of readability levels.

The Est. ORF score was computed using the results of a large-scale research study investigating the linkage between the Star Reading scores and estimates of oral reading fluency on a range of passages with grade-level-appropriate difficulty. An equipercentile linking was done between Star Reading scores and oral reading fluency, providing an estimate of the oral reading fluency for each Scaled Score unit in Star Reading for grades 1–4 independently. Results of the analysis can be found in "Additional Validation Evidence for Star Reading" on page 69. A table of selected Star Reading Scaled Scores and corresponding Est. ORF values can be found in "Appendix B: Detailed Evidence of Star Reading Validity" on page 139.

Instructional Reading Level (IRL)

The Instructional Reading Level is a criterion-referenced score that indicates the highest reading level at which the student can effectively be taught. In other words, IRLs tell you the reading level at which students can recognize words and comprehend written instructional material with some assistance. A sixth-grade student with an IRL of 4.0, for example, would be best served by instructional materials prepared at the fourth-grade level. IRLs are represented by either numbers or letters indicating a particular grade. Number codes represent IRLs for grades 1.0–12.9. IRL letter codes include PP (Pre-Primer), P (Primer, grades .1–.9), and PHS (Post-High School, grades 13.0+).

As a construct, instructional reading levels have existed in the field of reading education for over seventy years. During this time, a variety of assessment instruments have been developed using different measurement criteria that teachers can use to estimate IRL. Star Reading software determines IRL scores relative to 1995 updated vocabulary lists that are based on the Educational Development Laboratory's (EDL) *A Revised Core Vocabulary* (1969). 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 does not directly assess word recognition, Star Reading 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 IRL Scores

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

When Star Reading software determines that a student can answer 80 percent or more of the grade 13 items in the Star Reading test correctly, the student is assigned an IRL of Post-High School (PHS). This is the highest IRL that anyone can obtain when taking the Star Reading test.

Understanding IRL and GE Scores

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

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

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

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

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

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

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

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

Percentile Rank (PR)

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

The Percentile Rank scale is not an equal-interval scale. For example, for a student with a grade placement of 7.7, a Scaled Score of 1,119 corresponds

to a PR of 80, and a Scaled Score of 1,222 corresponds to a PR of 90. Thus, a difference of 103 Scaled Score points represents a 10-point difference in PR. However, for students at the same 7.7 grade placement, a Scaled Score of 843 corresponds to a PR of 50, and a Scaled Score of 917 corresponds to a PR of 60. While there is now only a 74-point difference in Scaled Scores, there is still a 10-point difference in PR. For this reason, PR scores should not be averaged or otherwise algebraically manipulated. NCE scores are much more appropriate for these activities.

Table 55 on page 119 and Table 56 on page 122 contain an abridged version of the Scaled Score to Percentile Rank conversion table that the Star Reading software uses. The actual table includes data for all of the monthly grade placement values from 1.0–12.9.

This table can be used to estimate PR values for tests that were taken when the grade placement value of a student was incorrect (see "Types of Test Scores" on page 99 for more information). If the error is caught right away, one always has the option of correcting the grade placement for the student and then having the student retest.

Normal Curve Equivalent (NCE)

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

Table 57 on page 126 provides the NCEs corresponding to integer PR values and facilitates the conversion of PRs to NCEs. Table 58 on page 127 provides the conversions from NCE to PR. The NCE values are given as a range of scores that convert to the corresponding PR value.

Student Growth Percentile (SGP)

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

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

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

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

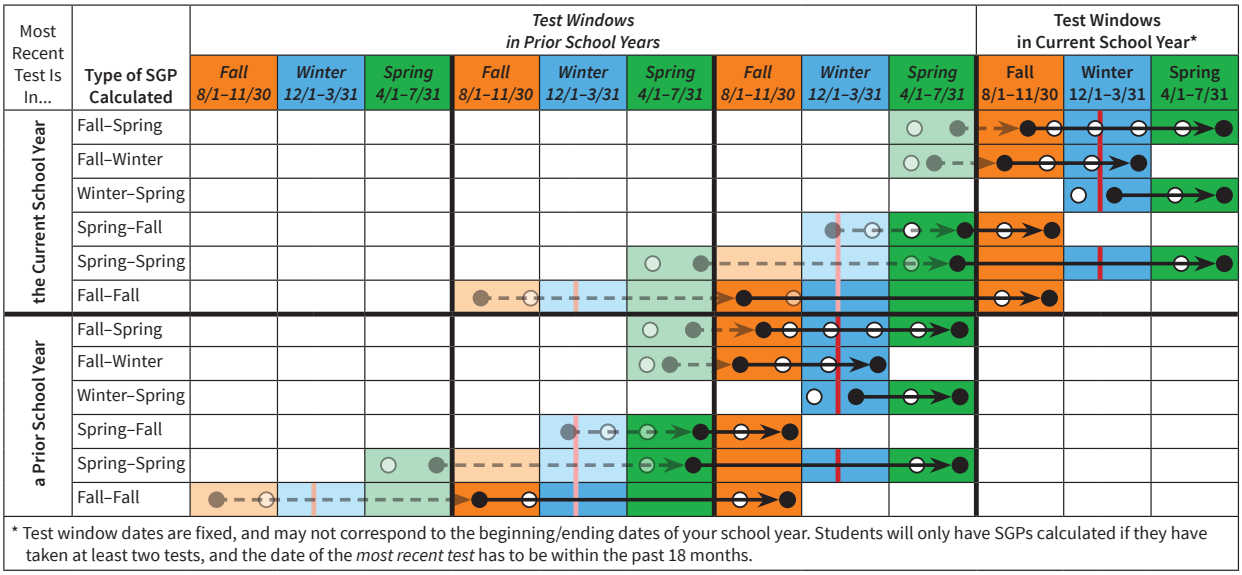
SGPs can be aggregated to describe typical growth for groups of students—for example, a class, grade, or school as a whole—by calculating the group's median, or middle, growth percentile. No matter how SGPs are aggregated, whether at the class, grade, or school level, the statistic and its interpretation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

1. We collect data for our growth norms during three different time periods: fall, winter, and spring. More information about these time periods is provided on page 111.

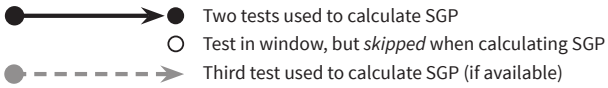
SGP is calculated for students who have taken at least two tests (a *current* test and a *prior* test) within at least two different testing windows (Fall, Winter, or Spring).

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

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



* Test window dates are fixed, and may not correspond to the beginning/ending dates of your school year. Students will only have SGPs calculated if they have taken at least two tests, and the date of the *most recent test* has to be within the past 18 months.



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

Lexile® Measures

In cooperation with MetaMetrics®, since August 2014, users of Star Reading have had the option of including Lexile measures on certain Star Reading score reports. Reported Lexile measures will range from BR400L to 1825L. (The "L" suffix identified the score as a Lexile measure. Where it appears, the

"BR" prefix indicates a score that is below 0 on the Lexile scale; such scores are typical of beginning readers.)

Lexile Measures of Students and Books: Measures of Student Reading Achievement and Text Readability

The ability to read and comprehend written text is important for academic success. Students may, however, benefit most from reading materials that match their reading ability/achievement: reading materials that are neither too easy nor too hard so as to maximize learning. To facilitate students' choices of appropriate reading materials, measures commonly referred to as *readability measures* are used in conjunction with students' reading achievement measures.

A text readability measure can be defined as a numeric scale, often derived analytically, that takes into account text characteristics that influence text comprehension or readability. An example of a readability measure is an age-level estimate of text difficulty. Among text characteristics that can affect text comprehension are sentence length and word difficulty.

A person's reading measure is a numeric score obtained from a reading achievement test, usually a standardized test such as Star Reading. A person's reading score quantifies his/her reading achievement level at a particular point in time.

Matching a student with text/books that target a student's interest and level of reading achievement is a two-step process: first, a student's reading achievement score is obtained by administering a standardized reading achievement test; second, the reading achievement score serves as an entry point into the readability measure to determine the difficulty level of text/books that would best support independent reading for the student. Optimally, a readability measure should match students with books that they are able to read and comprehend independently without boredom or frustration: books that are engaging yet slightly challenging to students based on the students' reading achievement and grade level.

Renaissance Learning's (RLI) readability measure is known as the Advantage/TASA Open Standard for Readability (ATOS). The *ATOS for Text* readability formula was developed through extensive research by RLI in conjunction with Touchstone Applied Science Associates, Inc. (TASA), now called Questar Assessment, Inc. A great many school libraries use ATOS book levels to index readability of their books. ATOS book levels, which are derived from *ATOS for Books* measures, express readability as grade levels; for example, an ATOS readability measure of 4.2 means that the book is at a difficulty level

appropriate for students reading at a level typical of students in the 4th grade, 2nd month. To match students to books at an appropriate level, the widely used Accelerated Reader system uses ATOS measures of readability and student's Grade Equivalent (GE) scores on standardized reading tests such as Star Reading.

Another widely-used system for matching readers to books at appropriate difficulty levels is The Lexile Framework® for Reading, developed by MetaMetrics, Inc. The Lexile scale is a common scale for both text measure (readability or text difficulty) and reader measure (reading achievement scores); in the Lexile Framework, both text difficulty and person reading ability are measured on the same scale. Unlike *ATOS for Books,* the Lexile Framework expresses a book's reading difficulty level (and students' reading ability levels) on a continuous scale ranging from below 0 to 1825 or more. Because some schools and school libraries use the Lexile Framework to index the reading difficulty levels of their books, there was a need to provide users of Star Reading with a student reading ability score compatible with the Lexile Framework.

In 2014, Metametrics, Inc., developed a means to translate Star Reading scale scores into equivalent Lexile measures of student reading ability. To do so, more than 200 MetaMetrics reading test items that had already been calibrated on the Lexile scale were administered in small numbers as unscored scale anchor items at the end of Star Reading tests. More than 250,000 students in grades 1 through 12 took up to 6 of those items as part of their Star Reading tests in April 2014. MetaMetrics' analysis of the Star Reading and Lexile anchor item response data yielded a means of transforming Star Reading's underlying Rasch scores into equivalent Lexile scores. That transformation, in turn, was used to develop a concordance table listing the Lexile equivalent of each unique Star Reading scale score.

In some cases, a range of text/book reading difficulty in which a student can read independently or with minimal guidance is desired. At Renaissance, we define the range of reading difficulty level that is neither too hard nor too easy as the Zone of Proximal Development (ZPD). The ZPD range allows, potentially, optimal learning to occur because students are engaged and appropriately challenged by reading materials that match their reading achievement and interest. The ZPD range is simply an approximation of the range of reading materials that is likely to benefit the student most. ZPD ranges are not absolute and teachers should also use their objective judgment to help students select reading books that enhance learning.

In a separate linking procedure, MetaMetrics compared the ATOS readability measures of thousands of books to the Lexile measures of the same books.

Analysis of those data yielded a table of equivalence between ATOS reading grade levels and Lexile readability measures. That equivalence table supports matching students to books regardless of whether a book's readability is measured using the Renaissance Learning ATOS system or the Lexile Framework created by MetaMetrics. Additionally, it supports translating ATOS ZPD ranges into equivalent ZPD ranges expressed on the Lexile scale.

Special Star Reading Scores

Most of the scores provided by Star Reading software are common measures of reading performance. Star Reading software also determines the Zone of Proximal Development.

Zone of Proximal Development (ZPD)

The Zone of Proximal Development (ZPD) defines the readability 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. Star Reading software uses Grade Equivalents to derive a student's ZPD score. Specifically, it relates the Grade Equivalent estimate of a student's reading ability with the range of most appropriate readability levels to use for reading practice. Table 59 on page 128 shows the relationship between GEs and ZPD scores.

The Zone of Proximal Development is especially useful for students who use Accelerated Reader, which provides readability levels on over 180,000 trade books and textbooks. Renaissance Learning developed the ZPD ranges according to Vygotskian theory, based on an analysis of Accelerated Reader book reading data from 80,000 students in the 1996–1997 school year. 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).

Grade Placement

Star Reading software uses the student's grade placement—grade and month of the school year—when determining the norm-referenced scores. The values of PR and NCE are based not only on what scaled score the student achieved but also on the grade placement of the student at the time of the test (for example, a second-grader in the seventh month with a scaled score of 395 would have a PR of 65, while a third-grader in the seventh month with

the same scaled score would have a PR of 41). Thus, it is crucial that student records indicate the proper grade when students take a Star Reading test, and that any testing in July or August reflects the proper understanding of how Star Reading software deals with these months in determining grade placement.

Indicating the Appropriate Grade Placement

The numeric representation of a student's grade placement is based on the specific month and day in which he or she takes a test. Although teachers indicate a student's grade level using whole numbers, Star Reading software automatically adds fractional increments to that grade level based on the month and day of the test. (Note: Grade placements for pre-K students are negative numbers.) To determine the appropriate increment, Star Reading software considers the standard school year to run from September—June and assigns increment values of .0–.9 to these months. Table 53 on page 101 summarizes the increment values assigned to each month.

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

- If teachers will use the July and August test scores to evaluate the student's reading performance at the beginning of the year, educators must make sure the following school year is set as the current school year in the Renaissance program at the time they administer the summer tests. Grades are automatically increased by one level in each successive school year, so promoting students to the next grade is not necessary. In this case, the increment value for July and August is 0.00 because these months are at the beginning of the school year.
- If teachers will use the test scores to evaluate the student's reading performance at the end of the school year, they must make sure the school year that has just ended is set as the current school year in the Renaissance program at the time they administer the summer tests. In this case, the increment value for July and August is 0.99 because these months are at the end of the school year that has passed.

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

If a school follows the standard school calendar used in Star Reading software and does not test in the summer, assigning the appropriate grade

placements for students is relatively easy. However, if students will be tested in July or August—whether it is for a summer reading program or because the normal calendar extends into these months—grade placements become an extremely important issue.

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

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

Compensating for Incorrect Grade Placements

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

Conversion Tables

Table 54: Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score		Enterprise Scaled Score	
	Low	High	Low	High
0	600	698	0	54
0.1	699	708	55	57
0.2	709	718	58	60
0.3	719	728	61	63
0.4	729	738	64	66
0.5	739	748	67	69
0.6	749	757	70	71
0.7	758	767	72	74
0.8	768	776	75	77
0.9	777	785	78	80
1	786	794	81	83
1.1	795	802	84	86
1.2	803	811	87	90
1.3	812	819	91	96
1.4	820	827	97	104
1.5	828	835	105	115
1.6	836	843	116	134
1.7	844	851	135	150
1.8	852	859	151	165
1.9	860	866	166	179
2	867	874	180	194
2.1	875	881	195	208
2.2	882	888	209	221
2.3	889	895	222	235
2.4	896	902	236	248
2.5	903	908	249	260
2.6	909	915	261	274

Table 54: Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score		Enterprise Scaled Score	
	Low	High	Low	High
2.7	916	921	275	285
2.8	922	927	286	298
2.9	928	934	299	314
3	935	940	315	326
3.1	941	946	327	340
3.2	947	951	341	351
3.3	952	957	352	364
3.4	958	963	365	375
3.5	964	968	376	388
3.6	969	973	389	400
3.7	974	978	401	415
3.8	979	983	416	428
3.9	984	988	429	443
4	989	993	444	454
4.1	994	998	455	464
4.2	999	1003	465	476
4.3	1004	1007	477	488
4.4	1008	1012	489	501
4.5	1013	1016	502	513
4.6	1017	1020	514	524
4.7	1021	1024	525	536
4.8	1025	1028	537	551
4.9	1029	1032	552	560
5	1033	1036	561	573
5.1	1037	1040	574	586
5.2	1041	1043	587	596
5.3	1044	1047	597	611
5.4	1048	1050	612	621
5.5	1051	1054	622	636
5.6	1055	1057	637	647
5.7	1058	1060	648	659

Table 54: Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score		Enterprise Scaled Score	
	Low	High	Low	High
5.8	1061	1063	660	672
5.9	1064	1066	673	683
6	1067	1069	684	697
6.1	1070	1072	698	712
6.2	1073	1075	713	725
6.3	1076	1078	726	743
6.4	1079	1080	744	761
6.5	1081	1083	762	779
6.6	1084	1085	780	789
6.7	1086	1088	790	805
6.8	1089	1090	806	816
6.9	1091	1093	817	835
7	1094	1095	836	846
7.1	1096	1097	847	854
7.2	1098	1099	855	864
7.3	1100	1101	865	877
7.4	1102	1103	878	886
7.5	1104	1105	887	894
7.6	1106	1107	895	901
7.7	1108	1109	902	907
7.8	1110	1111	908	912
7.9	1112	1112	913	916
8	1113	1114	917	924
8.1	1115	1116	925	936
8.2	1117	1117	937	941
8.3	1118	1119	942	951
8.4	1120	1120	952	957
8.5	1121	1122	958	967
8.6	1123	1123	968	971
8.7	1124	1125	972	980
8.8	1126	1126	981	985

Table 54: Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score		Enterprise Scaled Score	
	Low	High	Low	High
8.9	1127	1127	986	993
9	1128	1129	994	1014
9.1	1130	1130	1015	1025
9.2	1131	1131	1026	1035
9.3	1132	1132	1036	1042
9.4	1133	1134	1043	1058
9.5	1135	1135	1059	1065
9.6	1136	1136	1066	1073
9.7	1137	1137	1074	1086
9.8	1138	1138	1087	1096
9.9	1139	1139	1097	1101
10	1140	1140	1102	1106
10.1	1141	1141	1107	1115
10.2	1142	1142	1116	1123
10.3	1143	1143	1124	1132
10.4	1144	1144	1133	1140
10.5	1145	1145	1141	1148
10.6	1146	1146	1149	1154
10.7	1147	1147	1155	1161
10.8	1148	1148	1162	1166
10.9	1149	1149	1167	1171
11	1150	1150	1172	1175
11.1	1151	1151	1176	1180
11.2	1152	1152	1181	1185
11.3	1153	1153	1186	1190
11.4	1154	1154	1191	1197
11.5	1155	1155	1198	1203
11.6	1156	1156	1204	1208
11.7	1157	1157	1209	1213
11.8	1158	1158	1214	1216
11.9	1159	1159	1217	1219

Table 54: Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score		Enterprise Scaled Score	
	Low	High	Low	High
12	1160	1160	1220	1223
12.1	1161	1161	1224	1227
12.2	1162	1162	1228	1231
12.3	1163	1163	1232	1236
12.4	1164	1164	1237	1243
12.5	1165	1165	1244	1249
12.6	1166	1166	1250	1252
12.7	1167	1167	1253	1255
12.8	1168	1169	1256	1265
12.9	1170	1170	1266	1271
13	1171	1400	1272	1400

Table 55: Enterprise Scaled Score to Percentile Rank Conversions[a]

	Grade Placement												
PR	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	—	0	0	0	0	0	0	0	0	0	0	0	0
2	—	16	68	85	128	198	257	299	341	369	399	406	401
3	—	24	71	91	164	232	294	343	378	421	459	463	465
4	—	30	74	100	188	257	323	372	416	459	497	502	514
5	0	35	76	108	207	277	347	394	441	485	531	534	548
6	2	39	77	126	224	292	367	416	461	514	559	564	577
7	4	42	79	143	238	308	378	435	479	534	584	587	601
8	7	45	81	155	249	321	397	451	497	559	612	612	630
9	9	48	82	166	261	334	413	463	514	577	633	630	656
10	12	51	84	176	269	345	424	472	528	590	648	648	680
11	14	53	85	186	277	355	435	482	541	606	660	669	703
12	16	55	87	192	282	363	446	494	555	615	673	680	722
13	18	56	88	198	290	371	453	502	561	630	684	698	744
14	19	57	89	203	296	376	461	511	574	640	698	713	769
15	21	58	90	209	301	383	467	522	584	651	713	726	790
16	23	59	91	217	308	392	474	528	593	669	726	744	806
17	25	—	94	222	315	399	482	537	606	680	744	769	830
18	26	60	96	228	319	406	489	548	615	693	762	785	847
19	27	—	98	232	325	416	497	555	626	708	780	800	865
20	28	61	100	236	332	421	505	561	633	717	790	811	878
21	30	62	103	241	337	429	511	567	644	731	800	830	891
22	—	—	105	245	341	435	519	577	651	744	817	842	899
23	32	63	107	251	347	444	525	584	664	762	830	851	908
24	33	—	109	255	352	449	531	593	673	780	842	865	913
25	35	64	114	261	357	453	537	601	680	790	855	878	921
26	36	—	119	265	362	457	545	609	689	800	865	891	937
27	37	65	126	269	365	461	552	615	698	811	878	899	947
28	38	—	131	273	369	465	557	622	713	824	887	908	958
29	40	66	135	279	372	469	561	630	722	836	895	913	968
30	41	—	139	282	376	474	567	637	731	847	902	921	975

a. Each entry is the lowest Scaled Score for that grade and percentile.

Table 55: Enterprise Scaled Score to Percentile Rank Conversions[a]

	Grade Placement												
PR	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
31	42	—	145	286	380	479	574	644	744	855	908	931	986
32	43	67	149	290	386	485	577	651	769	865	913	942	1004
33	44	—	153	294	392	492	584	660	780	878	921	952	1026
34	45	68	157	299	397	497	590	669	790	887	931	964	1043
35	46	—	162	304	399	499	597	676	800	895	942	968	1059
36	47	—	166	308	404	505	606	680	811	902	947	975	1074
37	—	69	170	311	409	511	609	689	824	908	958	986	1097
38	48	—	174	315	416	514	615	698	830	910	968	994	1107
39	—	70	178	319	418	519	619	708	842	917	972	1015	1124
40	50	—	182	323	424	525	626	713	847	925	981	1026	1141
41	51	—	186	327	429	528	630	722	855	937	986	1043	1149
42	—	71	190	330	435	534	637	731	865	942	1004	1059	1162
43	52	—	193	334	439	537	640	744	878	952	1015	1074	1172
44	53	72	198	339	444	545	648	762	882	958	1036	1097	1181
45	54	—	199	343	446	548	651	769	891	968	1051	1107	1186
46	—	—	203	345	451	555	660	780	899	975	1066	1116	1198
47	55	73	207	350	455	557	664	790	905	981	1087	1133	1204
48	—	—	211	355	459	561	673	795	910	994	1102	1149	1214
49	56	74	215	357	461	564	680	806	917	1004	1107	1162	1217
50	—	—	219	362	465	570	684	811	921	1026	1124	1172	1224
51	57	75	222	365	469	577	693	824	931	1036	1141	1176	1228
52	—	—	226	367	472	581	703	830	942	1051	1149	1186	1237
53	—	76	230	371	477	587	708	842	947	1066	1162	1198	1244
54	58	—	232	374	479	590	717	847	958	1087	1172	1204	1253
55	—	77	236	376	485	597	726	855	968	1102	1176	1214	1256
56	59	78	239	380	489	601	731	865	972	1116	1186	1220	1266
57	—	—	243	383	494	609	744	878	981	1133	1191	1228	1279
58	—	79	247	389	499	612	762	887	994	1149	1204	1232	1285
59	60	80	251	394	502	619	769	891	1004	1162	1209	1244	1293
60	—	81	255	397	505	622	780	899	1026	1172	1217	1253	1298

a. Each entry is the lowest Scaled Score for that grade and percentile.

Table 55: Enterprise Scaled Score to Percentile Rank Conversions[a]

	Grade Placement												
PR	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
61	—	—	259	401	511	626	785	905	1036	1181	1224	1256	1301
62	61	82	263	404	514	633	795	910	1051	1186	1232	1266	1307
63	—	83	267	409	519	637	806	913	1066	1198	1237	1272	1309
64	62	84	271	416	522	644	811	921	1074	1209	1250	1285	1314
65	—	—	275	421	528	648	824	931	1102	1217	1256	1293	1315
66	63	85	279	424	531	656	830	942	1107	1224	1266	1298	1318
67	—	86	282	429	537	664	842	947	1133	1232	1279	1301	1319
68	64	87	288	435	541	669	851	958	1149	1237	1290	1307	1321
69	—	—	292	441	548	676	859	968	1162	1250	1296	1312	1323
70	—	88	296	446	555	684	865	975	1167	1256	1298	1315	1324
71	65	89	301	451	557	689	878	981	1176	1261	1304	1318	—
72	66	90	306	455	561	698	887	994	1186	1272	1309	1319	1327
73	—	91	311	459	564	708	895	1015	1198	1285	1312	1321	1329
74	—	94	315	463	570	717	902	1036	1209	1293	1315	1323	1330
75	67	97	319	467	577	722	908	1051	1214	1298	1318	1325	1332
76	—	99	323	472	584	731	910	1066	1220	1307	1320	1328	1334
77	68	102	330	477	590	744	917	1097	1228	1312	1321	1330	1336
78	—	105	334	482	593	762	925	1107	1244	1315	1324	1331	1338
79	69	107	339	489	601	775	937	1133	1250	1319	—	1333	1341
80	70	114	345	494	609	785	947	1149	1261	1321	1327	1335	1342
81	—	121	350	499	615	795	952	1162	1266	1322	1329	1336	—
82	71	131	357	505	622	806	964	1172	1279	1325	1330	1338	—
83	72	139	362	511	630	817	972	1186	1293	1327	1332	1339	1343
84	—	147	367	517	637	830	981	1198	1298	1329	1334	1340	—
85	73	153	372	525	644	842	994	1209	1307	1330	1336	1342	1344
86	74	161	378	534	651	855	1015	1220	1314	1334	1339	—	—
87	75	168	386	545	664	871	1036	1232	1318	1337	1341	—	—
88	76	176	394	552	676	887	1059	1250	1322	1340	1342	1343	—
89	77	184	401	559	689	899	1097	1261	—	—	1343	1344	1345
90	79	193	413	567	703	908	1116	1279	1327	1343	1344	—	—

a. Each entry is the lowest Scaled Score for that grade and percentile.

Table 55: Enterprise Scaled Score to Percentile Rank Conversions[a]

	Grade Placement												
PR	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
91	81	203	424	577	717	917	1141	1290	1330	1344	—	—	—
92	82	215	435	587	731	931	1167	1296	1332	—	—	1345	—
93	84	228	446	597	762	947	1181	1304	1334	1345	1345	—	1346
94	86	241	457	612	785	964	1204	1312	1338	—	—	—	—
95	88	257	469	630	811	981	1217	1318	1342	—	—	1346	—
96	91	277	489	648	847	1015	1232	1322	1343	1346	1346	—	—
97	101	299	508	680	882	1074	1256	1328	1345	—	—	—	1347
98	126	332	537	722	913	1162	1293	1334	1346	—	1347	1347	1348
99	174	389	587	817	972	1224	1318	1343	1347	1347	1348	1348	1350

a. Each entry is the lowest Scaled Score for that grade and percentile.

Table 56: Unified Scaled Score to Percentile Rank Conversions[a]

	Grade Placement												
PR	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	-	600	600	600	600	600	600	600	600	600	600	600	600
2	-	629	743	797	841	876	907	926	947	963	976	972	973
3	-	643	754	812	859	894	926	946	966	981	997	994	995
4	-	654	763	823	871	907	939	960	979	996	1012	1009	1013
5	600	663	770	832	881	917	950	971	989	1007	1023	1021	1027
6	602	670	776	840	890	925	958	979	998	1016	1032	1032	1036
7	607	676	781	847	897	932	966	985	1006	1024	1040	1040	1045
8	613	681	786	854	903	938	972	992	1012	1031	1047	1046	1053
9	617	686	790	860	909	944	978	998	1018	1038	1053	1053	1060
10	622	691	794	865	913	948	982	1002	1022	1042	1057	1058	1066
11	625	695	798	869	917	953	986	1006	1026	1046	1061	1063	1070
12	629	698	801	873	920	957	990	1009	1030	1050	1064	1066	1075
13	633	701	805	876	924	961	993	1013	1034	1054	1067	1069	1079
14	635	704	808	879	927	964	997	1016	1037	1057	1070	1073	1083
15	638	707	810	882	929	967	1000	1019	1040	1059	1073	1076	1086
16	642	710	813	886	932	970	1003	1022	1044	1062	1076	1080	1090

a. Each entry is the lowest Scaled Score for that grade and percentile.

Table 56: Unified Scaled Score to Percentile Rank Conversions[a]

	Grade Placement												
PR	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
17	645	713	816	889	935	973	1006	1025	1047	1065	1078	1083	1093
18	647	715	818	892	937	976	1008	1027	1049	1068	1081	1085	1096
19	649	717	821	894	940	979	1011	1030	1052	1071	1084	1088	1099
20	651	720	823	896	943	981	1014	1032	1055	1074	1086	1091	1102
21	654	722	826	899	945	984	1016	1035	1057	1076	1088	1093	1105
22	657	724	829	901	947	986	1018	1037	1060	1079	1091	1096	1108
23	659	726	831	904	950	989	1021	1040	1062	1081	1093	1098	1110
24	660	728	833	906	952	991	1023	1042	1064	1084	1095	1100	1112
25	662	730	835	909	954	993	1025	1044	1067	1086	1098	1103	1115
26	665	732	837	911	956	995	1027	1047	1069	1088	1100	1105	1117
27	667	733	840	913	958	997	1029	1049	1071	1090	1102	1108	1119
28	669	735	842	915	960	999	1031	1051	1073	1092	1104	1110	1121
29	671	736	844	918	962	1001	1033	1053	1075	1094	1106	1112	1123
30	673	738	846	920	964	1004	1035	1055	1077	1096	1108	1114	1125
31	675	739	849	922	966	1005	1037	1057	1080	1098	1110	1116	1127
32	677	740	851	924	968	1007	1038	1059	1082	1100	1112	1118	1129
33	679	742	853	926	970	1009	1040	1060	1084	1101	1114	1120	1131
34	681	744	856	928	972	1011	1042	1062	1086	1103	1116	1122	1133
35	682	745	858	930	973	1012	1044	1064	1088	1105	1118	1123	1135
36	684	746	860	932	975	1014	1046	1066	1090	1107	1119	1125	1137
37	686	748	862	934	977	1016	1047	1068	1092	1109	1121	1127	1139
38	687	750	864	935	979	1017	1049	1070	1093	1111	1123	1128	1141
39	689	751	866	937	980	1019	1050	1071	1095	1113	1124	1130	1143
40	690	752	868	939	982	1021	1052	1073	1096	1115	1126	1132	1145
41	692	754	870	941	984	1022	1053	1075	1098	1116	1127	1133	1147
42	693	755	872	942	986	1024	1055	1077	1100	1118	1129	1135	1148
43	694	757	874	944	987	1025	1056	1078	1102	1120	1130	1137	1150
44	695	758	876	946	989	1027	1058	1080	1104	1121	1132	1139	1152
45	697	759	877	948	990	1028	1059	1082	1105	1123	1134	1141	1153
46	698	761	879	949	992	1030	1061	1084	1107	1124	1135	1142	1155

a. Each entry is the lowest Scaled Score for that grade and percentile.

Table 56: Unified Scaled Score to Percentile Rank Conversions[a]

	Grade Placement												
PR	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
47	699	762	881	951	994	1031	1062	1085	1109	1126	1138	1145	1156
48	700	764	883	953	996	1033	1064	1087	1111	1127	1140	1146	1158
49	702	765	885	954	997	1034	1066	1089	1112	1129	1141	1148	1160
50	703	767	887	956	999	1036	1067	1090	1114	1130	1143	1150	1161
51	704	769	889	958	1001	1038	1069	1092	1116	1132	1145	1152	1162
52	706	770	891	959	1002	1039	1070	1093	1118	1134	1146	1153	1164
53	707	772	893	961	1004	1041	1072	1095	1119	1136	1148	1155	1165
54	708	774	894	963	1006	1042	1074	1096	1121	1138	1150	1157	1167
55	710	776	896	964	1007	1044	1075	1098	1123	1140	1151	1158	1168
56	711	778	898	966	1008	1045	1077	1100	1124	1142	1153	1160	1170
57	713	780	900	967	1010	1047	1079	1102	1126	1144	1154	1162	1172
58	714	782	902	969	1012	1048	1081	1104	1127	1146	1156	1164	1173
59	715	784	904	971	1013	1050	1082	1105	1129	1148	1157	1165	1175
60	717	786	906	972	1014	1051	1084	1107	1130	1150	1159	1167	1177
61	718	788	908	974	1016	1052	1085	1109	1132	1152	1161	1168	1178
62	720	790	910	975	1017	1054	1087	1111	1134	1153	1163	1170	1179
63	721	792	912	977	1019	1055	1089	1112	1135	1155	1164	1172	1181
64	723	794	914	979	1020	1057	1090	1114	1137	1157	1166	1173	1182
65	725	796	916	981	1022	1058	1092	1116	1139	1158	1168	1175	1184
66	726	798	918	982	1023	1060	1093	1118	1141	1160	1170	1177	1186
67	728	800	920	984	1025	1062	1095	1119	1143	1162	1172	1178	1187
68	729	802	923	986	1026	1063	1097	1121	1145	1164	1174	1181	1189
69	731	804	925	988	1028	1065	1099	1123	1147	1166	1176	1182	1191
70	733	806	927	990	1030	1067	1100	1124	1149	1168	1177	1184	1192
71	734	808	929	992	1031	1068	1102	1126	1151	1170	1179	1186	1194
72	736	810	931	994	1033	1070	1104	1128	1153	1171	1181	1187	1195
73	738	813	933	996	1034	1072	1106	1129	1154	1173	1182	1189	1197
74	739	816	935	998	1036	1074	1108	1132	1156	1175	1184	1191	1199
75	741	819	937	1000	1038	1075	1109	1134	1158	1177	1186	1194	1201
76	743	822	939	1002	1040	1077	1111	1136	1160	1180	1188	1196	1203

a. Each entry is the lowest Scaled Score for that grade and percentile.

Table 56: Unified Scaled Score to Percentile Rank Conversions[a]

	Grade Placement												
PR	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
77	744	825	942	1004	1042	1079	1113	1139	1162	1182	1190	1198	1205
78	746	828	944	1006	1043	1081	1115	1141	1165	1184	1192	1200	1207
79	748	831	946	1008	1045	1083	1117	1143	1166	1187	1194	1202	1210
80	750	834	949	1010	1047	1085	1119	1146	1169	1189	1195	1204	1211
81	752	838	951	1012	1049	1087	1120	1148	1170	1190	1197	1205	1212
82	755	842	954	1014	1051	1089	1122	1150	1172	1193	1199	1207	1214
83	757	846	956	1016	1053	1091	1124	1152	1175	1195	1201	1208	1215
84	759	849	959	1018	1055	1093	1126	1155	1177	1197	1203	1209	1217
85	762	853	962	1021	1057	1095	1128	1157	1180	1200	1205	1211	1218
86	765	857	965	1024	1059	1098	1130	1160	1183	1203	1208	1213	1221
87	769	861	968	1027	1062	1101	1132	1163	1186	1206	1210	1214	1222
88	772	865	971	1029	1065	1104	1135	1166	1189	1210	1212	1216	1224
89	776	869	974	1032	1068	1107	1139	1169	1193	1215	1215	1217	1225
90	781	874	978	1035	1071	1110	1142	1172	1195	1216	1218	1220	1228
91	785	879	982	1038	1074	1113	1145	1174	1198	1219	1221	1222	1231
92	790	885	986	1041	1077	1116	1149	1176	1201	1223	1223	1225	1234
93	795	892	990	1044	1081	1119	1152	1179	1203	1226	1226	1229	1237
94	799	899	995	1048	1085	1122	1156	1182	1206	1227	1231	1232	1242
95	805	907	1001	1053	1090	1126	1159	1186	1211	1231	1234	1237	1246
96	813	917	1008	1058	1096	1130	1163	1190	1214	1237	1238	1242	1252
97	824	929	1015	1066	1103	1137	1168	1196	1221	1241	1245	1250	1260
98	840	943	1025	1075	1112	1148	1175	1203	1226	1249	1251	1256	1268
99	865	969	1041	1091	1124	1161	1186	1215	1239	1261	1267	1273	1285

a. Each entry is the lowest Scaled Score for that grade and percentile.

Table 57: Percentile Rank to Normal Curve Equivalent Conversions

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

Table 58: Normal Curve Equivalent to Percentile Rank Conversion

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

Table 59: Grade Equivalent to ZPD Conversions

	ZPD Range			ZPD Range			ZPD Range	
GE	Low	High	GE	Low	High	GE	Low	High
0.0	0.0	1.0	4.4	3.2	4.9	8.8	4.6	8.8
0.1	0.1	1.1	4.5	3.2	5.0	8.9	4.6	8.9
0.2	0.2	1.2	4.6	3.2	5.1	9.0	4.6	9.0
0.3	0.3	1.3	4.7	3.3	5.2	9.1	4.6	9.1
0.4	0.4	1.4	4.8	3.3	5.2	9.2	4.6	9.2
0.5	0.5	1.5	4.9	3.4	5.3	9.3	4.6	9.3
0.6	0.6	1.6	5.0	3.4	5.4	9.4	4.6	9.4
0.7	0.7	1.7	5.1	3.5	5.5	9.5	4.7	9.5
0.8	0.8	1.8	5.2	3.5	5.5	9.6	4.7	9.6
0.9	0.9	1.9	5.3	3.6	5.6	9.7	4.7	9.7
1.0	1.0	2.0	5.4	3.6	5.6	9.8	4.7	9.8
1.1	1.1	2.1	5.5	3.7	5.7	9.9	4.7	9.9
1.2	1.2	2.2	5.6	3.8	5.8	10.0	4.7	10.0
1.3	1.3	2.3	5.7	3.8	5.9	10.1	4.7	10.1
1.4	1.4	2.4	5.8	3.9	5.9	10.2	4.7	10.2
1.5	1.5	2.5	5.9	3.9	6.0	10.3	4.7	10.3
1.6	1.6	2.6	6.0	4.0	6.1	10.4	4.7	10.4
1.7	1.7	2.7	6.1	4.0	6.2	10.5	4.8	10.5
1.8	1.8	2.8	6.2	4.1	6.3	10.6	4.8	10.6
1.9	1.9	2.9	6.3	4.1	6.3	10.7	4.8	10.7
2.0	2.0	3.0	6.4	4.2	6.4	10.8	4.8	10.8
2.1	2.1	3.1	6.5	4.2	6.5	10.9	4.8	10.9
2.2	2.1	3.1	6.6	4.2	6.6	11.0	4.8	11.0
2.3	2.2	3.2	6.7	4.2	6.7	11.1	4.8	11.1
2.4	2.2	3.2	6.8	4.3	6.8	11.2	4.8	11.2
2.5	2.3	3.3	6.9	4.3	6.9	11.3	4.8	11.3
2.6	2.4	3.4	7.0	4.3	7.0	11.4	4.8	11.4
2.7	2.4	3.4	7.1	4.3	7.1	11.5	4.9	11.5
2.8	2.5	3.5	7.2	4.3	7.2	11.6	4.9	11.6
2.9	2.5	3.5	7.3	4.4	7.3	11.7	4.9	11.7
3.0	2.6	3.6	7.4	4.4	7.4	11.8	4.9	11.8
3.1	2.6	3.7	7.5	4.4	7.5	11.9	4.9	11.9

Table 59: Grade Equivalent to ZPD Conversions

	ZPD Range			ZPD Range			ZPD Range	
GE	Low	High	GE	Low	High	GE	Low	High
3.2	2.7	3.8	7.6	4.4	7.6	12.0	4.9	12.0
3.3	2.7	3.8	7.7	4.4	7.7	12.1	4.9	12.1
3.4	2.8	3.9	7.8	4.5	7.8	12.2	4.9	12.2
3.5	2.8	4.0	7.9	4.5	7.9	12.3	4.9	12.3
3.6	2.8	4.1	8.0	4.5	8.0	12.4	4.9	12.4
3.7	2.9	4.2	8.1	4.5	8.1	12.5	5.0	12.5
3.8	2.9	4.3	8.2	4.5	8.2	12.6	5.0	12.6
3.9	3.0	4.4	8.3	4.5	8.3	12.7	5.0	12.7
4.0	3.0	4.5	8.4	4.5	8.4	12.8	5.0	12.8
4.1	3.0	4.6	8.5	4.6	8.5	12.9	5.0	12.9
4.2	3.1	4.7	8.6	4.6	8.6	13.0	5.0	13.0
4.3	3.1	4.8	8.7	4.6	8.7			

Table 60: Scaled Score to Instructional Reading Level Conversions[a]

	Unified Scaled Score		Enterprise Scaled Score	
IRL	**Low**	**High**	**Low**	**High**
Pre-Primer (PP): < 0	600	838	0	124
Primer (P): 0.1–0.9	839	855	125	159
1	856	860	160	168
1.1	861	864	169	176
1.2	865	869	177	185
1.3	870	873	186	194
1.4	874	878	195	203
1.5	879	882	204	212
1.6	883	887	213	220
1.7	888	892	221	229
1.8	893	896	230	238
1.9	897	901	239	247
2	902	905	248	256
2.1	906	910	257	266
2.2	911	915	267	275
2.3	916	920	276	284
2.4	921	924	285	293
2.5	925	929	294	304
2.6	930	934	305	315
2.7	935	939	316	325
2.8	940	944	326	336
2.9	945	948	337	346
3	949	954	347	359
3.1	955	959	360	369
3.2	960	964	370	379
3.3	965	970	380	394
3.4	971	975	395	407
3.5	976	980	408	423
3.6	981	986	424	439
3.7	987	991	440	451
3.8	992	997	452	462

a. The figures in this table only apply to individual students, not groups.

Table 60: Scaled Score to Instructional Reading Level Conversions[a]

	Unified Scaled Score		Enterprise Scaled Score	
IRL	**Low**	**High**	**Low**	**High**
3.9	998	1002	463	474
4	1003	1006	475	487
4.1	1007	1011	488	498
4.2	1012	1015	499	512
4.3	1016	1019	513	523
4.4	1020	1024	524	537
4.5	1025	1028	538	553
4.6	1029	1032	554	563
4.7	1033	1037	564	577
4.8	1038	1041	578	590
4.9	1042	1045	591	607
5	1046	1048	608	616
5.1	1049	1050	617	624
5.2	1051	1053	625	633
5.3	1054	1055	634	642
5.4	1056	1058	643	652
5.5	1059	1060	653	662
5.6	1061	1063	663	673
5.7	1064	1065	674	682
5.8	1066	1068	683	694
5.9	1069	1070	695	706
6	1071	1074	707	725
6.1	1075	1079	726	752
6.2	1080	1083	753	780
6.3	1084	1087	781	801
6.4	1088	1091	802	826
6.5	1092	1095	827	848
6.6	1096	1099	849	868
6.7	1100	1103	869	890
6.8	1104	1107	891	904
6.9	1108	1112	905	916

a. The figures in this table only apply to individual students, not groups.

Table 60: Scaled Score to Instructional Reading Level Conversions[a]

	Unified Scaled Score		Enterprise Scaled Score	
IRL	**Low**	**High**	**Low**	**High**
7	1113	1112	917	918
7.1	1113	1113	919	920
7.2	1114	1113	921	922
7.3	1114	1114	923	924
7.4	1115	1114	925	928
7.5	1115	1115	929	930
7.6	1116	1115	931	934
7.7	1116	1116	935	937
7.8	1117	1116	938	939
7.9	1117	1117	940	942
8	1118	1118	943	948
8.1	1119	1119	949	954
8.2	1120	1120	955	960
8.3	1121	1121	961	966
8.4	1122	1122	967	970
8.5	1123	1124	971	974
8.6	1125	1125	975	981
8.7	1126	1126	982	988
8.8	1127	1127	989	998
8.9	1128	1128	999	1011
9	1129	1129	1012	1022
9.1	1130	1130	1023	1034
9.2	1131	1131	1035	1042
9.3	1132	1133	1043	1050
9.4	1134	1134	1051	1058
9.5	1135	1135	1059	1067
9.6	1136	1136	1068	1076
9.7	1137	1137	1077	1090
9.8	1138	1138	1091	1098
9.9	1139	1139	1099	1104
10	1140	1140	1105	1111

a. The figures in this table only apply to individual students, not groups.

Table 60: Scaled Score to Instructional Reading Level Conversions[a]

	Unified Scaled Score		Enterprise Scaled Score	
IRL	**Low**	**High**	**Low**	**High**
10.1	1141	1141	1112	1121
10.2	1142	1142	1122	1130
10.3	1143	1143	1131	1139
10.4	1144	1144	1140	1147
10.5	1145	1146	1148	1155
10.6	1147	1147	1156	1161
10.7	1148	1148	1162	1167
10.8	1149	1149	1168	1172
10.9	1150	1150	1173	1177
11	1151	1154	1178	1203
11.1	1155	1159	1204	1221
11.2	1160	1164	1222	1243
11.3	1165	1168	1244	1264
11.4	1169	1173	1265	1290
11.5	1174	1177	1291	1303
11.6	1178	1182	1304	1314
11.7	1183	1187	1315	1319
11.8	1188	1191	1320	1324
11.9	1192	1196	1325	1328
12	1197	1198	1329	1330
12.1	1199	1201	1331	1332
12.2	1202	1203	1333	1335
12.3	1204	1206	1336	1337
12.4	1207	1208	1338	1340
12.5	1209	1211	1341	1341
12.6	1212	1213	1342	1342
12.7	1214	1216	1343	1343
12.8	1217	1218	1344	1344
12.9	1219	1221	1345	1345
Post-High School (PHS)	1222	1400	1346	1400

a. The figures in this table only apply to individual students, not groups.

Table 61: Relating Star Early Literacy Enterprise Scores to Star Reading Enterprise Scores

Star Early Literacy			Star Reading				
Enterprise Scaled Score Range		Literacy Classification	Enterprise Scaled Score Range		GE	ZPD Range	Recommended Assessment(s)
Low	High		Low	High			
300	382	Emergent Reader	NA	NA	NA	NA	Star Early Literacy Enterprise
383	513		0	54	0	0–1	
514	531		55	57	0.1	0.1–1.1	
532	549		58	60	0.2	0.2–1.2	
550	568		61	63	0.3	0.3–1.3	
569	586		64	66	0.4	0.4–1.4	
587	605		67	69	0.5	0.5–1.5	
606	622		70	71	0.6	0.6–1.6	
623	641		72	74	0.7	0.7–1.7	
642	657		75	77	0.8	0.8–1.8	
658	673		78	80	0.9	0.9–1.9	
674	689	Transitional Reader SEL SS = 675	81	83	1	1–2	
690	703		84	86	1.1	1.1–2.1	
704	717		87	90	1.2	1.2–2.2	
718	730		91	96	1.3	1.3–2.3	
731	742		97	104	1.4	1.4–2.4	
743	753		105	115	1.5	1.5–2.5	Star Early Literacy Enterprise and Star Reading
754	764		116	134	1.6	1.6–2.6	
765	774		135	150	1.7	1.7–2.7	
775	784	Probable Reader SEL SS = 775	151	165	1.8	1.8–2.8	
785	792		166	179	1.9	1.9–2.9	Star Reading
793	800		180	194	2.0	2.0–3.0	
801	807		195	208	2.1	2.1–3.1	
808	814		209	221	2.2	2.1–3.1	
815	820		222	235	2.3	2.2–3.2	
821	826		236	248	2.4	2.2–3.2	
827	831		249	260	2.5	2.3–3.3	
832	836		261	274	2.6	2.4–3.4	
837	840		275	285	2.7	2.4–3.4	

Table 61: Relating Star Early Literacy Enterprise Scores to Star Reading Enterprise Scores

Star Early Literacy			Star Reading				
Enterprise Scaled Score Range		Literacy Classification	Enterprise Scaled Score Range		GE	ZPD Range	Recommended Assessment(s)
Low	High		Low	High			
841	844	Probable Reader SEL SS = 775 (continued)	286	298	2.8	2.5–3.5	Star Reading (continued)
845	848		299	314	2.9	2.5–3.5	
849	852		315	326	3.0	2.6–3.6	
853	855		327	340	3.1	2.6–3.7	
856	857		341	351	3.2	2.7–3.8	
858	860		352	364	3.3	2.7–3.8	
861	863		365	375	3.4	2.8–3.9	
864	865		376	388	3.5	2.8–4.0	
866	867		389	400	3.6	2.8–4.1	
868	869		401	415	3.7	2.9–4.2	
870	870		416	428	3.8	2.9–4.3	
871	900		429	443	3.9	3.0–4.4	

Table 62: Relating Star Early Literacy Scores to Star Reading Scores on the Unified Scale

Star Early Literacy			Star Reading		
Literacy Classification	Unified Scaled Score Range		GE	ZPD Range	Recommended Assessment(s)
	Low	High			
Emergent Reader	NA	NA	NA	NA	Star Early Literacy Enterprise
	600	698	0	0–1	
	699	708	0.1	0.1–1.1	
	709	718	0.2	0.2–1.2	
	719	728	0.3	0.3–1.3	
	729	738	0.4	0.4–1.4	
	739	748	0.5	0.5–1.5	
	749	757	0.6	0.6–1.6	
	758	767	0.7	0.7–1.7	
	768	776	0.8	0.8–1.8	
	777	785	0.9	0.9–1.9	
Transitional Reader SEL SS = 786	786	794	1	1–2	
	795	802	1.1	1.1–2.1	
	803	811	1.2	1.2–2.2	
	812	819	1.3	1.3–2.3	
	820	827	1.4	1.4–2.4	
	828	835	1.5	1.5–2.5	Star Early Literacy Enterprise and Star Reading
	836	843	1.6	1.6–2.6	
	844	851	1.7	1.7–2.7	
Probable Reader SEL SS = 852	852	859	1.8	1.8–2.8	
	860	866	1.9	1.9–2.9	Star Reading
	867	874	2.0	2.0–3.0	
	875	881	2.1	2.1–3.1	
	882	888	2.2	2.1–3.1	
	889	895	2.3	2.2–3.2	
	896	902	2.4	2.2–3.2	
	903	908	2.5	2.3–3.3	
	909	915	2.6	2.4–3.4	
	916	921	2.7	2.4–3.4	

Table 62: Relating Star Early Literacy Scores to Star Reading Scores on the Unified Scale

Star Early Literacy			Star Reading		
Literacy Classification	Unified Scaled Score Range		GE	ZPD Range	Recommended Assessment(s)
	Low	High			
Probable Reader (continued)	922	927	2.8	2.5–3.5	Star Reading (continued)
	928	934	2.9	2.5–3.5	
	935	940	3.0	2.6–3.6	
	941	946	3.1	2.6–3.7	
	947	951	3.2	2.7–3.8	
	952	957	3.3	2.7–3.8	
	958	963	3.4	2.8–3.9	
	964	968	3.5	2.8–4.0	
	969	973	3.6	2.8–4.1	
	974	978	3.7	2.9–4.2	
	979	983	3.8	2.9–4.3	
	984	988	3.9	3.0–4.4	

Appendix A: Estimated Oral Reading Fluency

Table 63: Estimated Oral Reading Fluency (Est. ORF) Given in Words Correct per Minute (WCPM) by Grade for Selected Star Reading Scale Score Units (SR SS)

	Grade			
SR SS	**1**	**2**	**3**	**4**
50	0	4	0	8
100	29	30	32	31
150	41	40	43	41
200	55	52	52	47
250	68	64	60	57
300	82	78	71	69
350	92	92	80	80
400	111	106	97	93
450	142	118	108	104
500	142	132	120	115
550	142	152	133	127
600	142	175	147	137
650	142	175	157	145
700	142	175	167	154
750	142	175	170	168
800	142	175	170	184
850–1400	142	175	170	190

Appendix B: Detailed Evidence of Star Reading Validity

The Validity chapter of this technical manual places its emphasis on summaries of Star Reading validity evidence, and on recent evidence which comes primarily from the 34-item, standards-based version of the assessment, which was introduced in 2011. However, the abundance of earlier evidence, and evidence related to the 25-item Star Reading versions, is all part of the accumulation of technical support for the validity and usefulness of Star Reading. Much of that cumulative evidence is presented in this appendix, to ensure that the historical continuity of research in support of Star Reading validity is not lost. The material that follows touches on the following list of topics:

- Relationship of Star Reading Scores to Scores on Other Tests of Reading Achievement
- Relationship of Star Reading Scores to Scores on State Tests of Accountability in Reading
- Relationship of Star Reading Enterprise Scores to Scores on Previous Versions
- Data from Post-Publication Studies
- Linking Star and State Assessments: Comparing Student- and School-Level Data
- Classification Accuracy and Screening Data Reported to The National Center on Response to Intervention (NCRTI)

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

During the Star Reading 2.0 norming study, schools submitted data on how their students performed on several standardized tests of reading achievement as well as their students' Star Reading results. This data included test results for more than 12,000 students from such tests as the California Achievement Test (CAT), the Comprehensive Test of Basic Skills (CTBS), the Iowa Test of Basic Skills (ITBS), the Metropolitan Achievement Test (MAT), the Stanford Achievement Test (SAT9), and several statewide tests.

Computing the correlation coefficients was a two-step process. First, where necessary, data were placed onto a common scale. If Scaled Scores were available, they could be correlated with Star Reading 2.0 Scaled Scores. However, since Percentile Ranks (PRs) are not on an equal-interval scale, when PRs were reported for the other tests, they were converted into Normal Curve Equivalents (NCEs). Scaled Scores or NCE scores were then used to compute the Pearson product-moment correlation coefficients.

In an ongoing effort to gather evidence for the validity of Star Reading scores, continual research on score validity has been undertaken. In addition to original validity data gathered at the time of initial development, numerous other studies have investigated the correlations between Star Reading tests and other external measures. In addition to gathering concurrent validity estimates, predictive validity estimates have also been investigated.

Table 64, Table 65, Table 66, and Table 67 present the correlation coefficients between the scores on the Star Reading test and each of the other tests for which data were received. Table 64 and Table 65 display "concurrent validity" data; that is, correlations between Star Reading test scores and other tests administered within a two-month time period. The date of administration ranged from spring 1999–spring 2013. More recently, data have become available for analyses regarding the predictive validity of Star Reading. Predictive validity provides an estimate of the extent to which scores on the Star Reading test predicted scores on criterion measures given at a later point in time, operationally defined as more than 2 months between the Star test (predictor) and the criterion test. Predictive validity provides an estimate of the linear relationship between Star scores and scores on tests covering a similar academic domain. Predictive correlations are attenuated by time due to the fact that students are gaining skills in the interim between testing occasions, and also by differences between the tests' content specifications. Table 66 and Table 67 present predictive validity coefficients.

The tables are presented in two parts. Table 64 and Table 66 display validity coefficients for grades 1–6, and Table 65 and Table 67 display the validity coefficients for grades 7–12. The bottom of each table presents a grade-by-grade summary, including the total number of students for whom test data were available, the number of validity coefficients for that grade, and the average value of the validity coefficients.

The within-grade average concurrent validity coefficients for grades 1–6 varied from 0.72–0.80, with an overall average of 0.74. The within-grade average concurrent validity for grades 7–12 ranged from 0.65–0.76, with an overall average of 0.72. Predictive validity coefficients ranged from 0.69–0.72 in grades 1–6, with an average of 0.71. In grades 7–12 the predictive validity

coefficients ranged from 0.72–0.87 with an average of 0.80. The other validity coefficient within-grade averages (for Star Reading 2.0 with external tests administered prior to spring 1999, Table 68 and Table 69) varied from 0.60–0.77; the overall average was 0.72. The process of establishing the validity of a test is laborious, and it usually takes a significant amount of time. As a result, the validation of the Star Reading 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 users who collect relevant data are encouraged to contact Renaissance Learning.

Since correlation coefficients are available for many different test editions, forms, and dates of administration, many of the tests have several validity coefficients associated with them. Data were omitted from the tabulations if (a) test data quality could not be verified or (b) when sample size was very small. In general, these correlation coefficients reflect very well on the validity of the Star Reading test as a tool for placement in Reading. In fact, the correlations are similar in magnitude to the validity coefficients of these measures with each other. These validity results, combined with the supporting evidence of reliability and minimization of SEM estimates for the Star Reading test, provide a quantitative demonstration of how well this innovative instrument in reading achievement assessment performs.

Table 64: Concurrent Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a]

			1		2		3		4		5		6	
Test Form	**Date**	**Score**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**
Arkansas Augmented Benchmark Examination (AABE)														
AABE	S 08	SS	–	–	–	–	2,858	0.78*	2,588	0.73*	1,897	0.73*	1,176	0.75*
AIMSweb														
R-CBM	S 12	correct	15	0.65*	72	0.28*	41	0.17	44	0.48*	–	–	–	–
California Achievement Test (CAT)														
CAT	S 99	SS	93	0.80*	36	0.67*	–	–	34	0.72*	146	0.76*	–	–
CAT/5	F 10–11	SS	68	0.79*	315	0.72*	410	0.69*	302	0.71*	258	0.71*	196	0.69*
Canadian Achievement Test (CAT)														
CAT/2	F 10–11		–	–	–	–	21	0.80*	31	0.84*	23	0.75*	–	–
Colorado Student Assessment Program (CSAP)														
CSAP	S 06	SS	–	–	–	–	82	0.75*	79	0.83*	93	0.68*	280	0.80*

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 64: Concurrent Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a]

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
Comprehensive Test of Basic Skills (CTBS)														
CTBS/4	S 99	NCE	–	–	–	–	–	–	18	0.81*	–	–	–	–
CTBS/A-19/20	S 99	SS	–	–	–	–	–	–	–	–	–	–	8	0.91*
Delaware Student Testing Program (DSTP) – Reading														
DSTP	S 05	SS	–	–	–	–	104	0.57*	–	–	–	–	–	–
DSTP	S 06	SS	–	–	158	0.68*	126	0.43*	141	0.62*	157	0.59*	75	0.66*
Dynamic Indicators of Basic Early Literacy Skills (DIBELS) – Oral Reading Fluency														
DIBELS	F 05	WCPM	–	–	59	0.78*	–	–	–	–	–	–	–	–
DIBELS	W 06	WCPM	61	0.87*	55	0.75*	–	–	–	–	–	–	–	–
DIBELS	S 06	WCPM	67	0.87*	63	0.71*	–	–	–	–	–	–	–	–
DIBELS	F 06	WCPM	–	–	515	0.78*	354	0.81*	202	0.72*	–	–	–	–
DIBELS	W 07	WCPM	208	0.75*	415	0.73*	175	0.69*	115	0.71*	–	–	–	–
DIBELS	S 07	WCPM	437	0.81*	528	0.70*	363	0.66*	208	0.54*	–	–	–	–
DIBELS	F 07	WCPM	–	–	626	0.79*	828	0.73*	503	0.73*	46	0.73*	–	–
Florida Comprehensive Assessment Test (FCAT)														
FCAT	S 06	SS	–	–	–	–	–	–	41	0.65*	–	–	–	–
FCAT	S 06–08	SS	–	–	–	–	10,169	0.76*	8,003	0.73*	5,474	0.73*	1,188	0.67*
Florida Comprehensive Assessment Test (FCAT 2.0)														
FCAT 2.0	S 13	SS	–	–	–	–	3,641	0.83*	3,025	0.84*	2,439	0.83*	145	0.81*
Gates–MacGinitie Reading Test (GMRT)														
GMRT/2nd Ed	S 99	NCE	–	–	21	0.89*	–	–	–	–	–	–	–	–
GMRT/L-3rd	S 99	NCE	–	–	127	0.80*	–	–	–	–	–	–	–	–
Idaho Standards Achievement Test (ISAT)														
ISAT	S 07–09	SS	–	–	–	–	3,724	0.75*	2,956	0.74*	2,485	0.74*	1,309	0.75*
Illinois Standards Achievement Test – Reading														
ISAT	S 05	SS	–	–	106	0.71*	594	0.76*	–	–	449	0.70*	–	–
ISAT	S 06	SS	–	–	–	–	140	0.80*	144	0.80*	146	0.72	–	–

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 64: Concurrent Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a]

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
Iowa Test of Basic Skills (ITBS)														
ITBS–Form K	S 99	NCE	40	0.75*	36	0.84*	26	0.82*	28	0.89*	79	0.74*	–	–
ITBS–Form L	S 99	NCE	–	–	–	–	18	0.70*	29	0.83*	41	0.78*	38	0.82*
ITBS–Form M	S 99	NCE	–	–	–	–	158	0.81*	–	–	125	0.84*	–	–
ITBS–Form K	S 99	SS	–	–	58	0.74*	–	–	54	0.79*	–	–	–	–
ITBS–Form L	S 99	SS	–	–	–	–	45	0.73*	–	–	–	–	50	0.82*
Kansas State Assessment Program (KSAP)														
KSAP	S 06–08	SS	–	–	–	–	4,834	0.61*	4,045	0.61*	3,332	0.63*	1,888	0.65*
Kentucky Core Content Test (KCCT)														
KCCT	S 08–10	SS	–	–	–	–	10,776	0.60*	8,885	0.56*	7,147	0.53*	5,003	0.57*
Metropolitan Achievement Test (MAT)														
MAT–7th Ed.	S 99	NCE	–	–	–	–	–	–	46	0.79*	–	–	–	–
MAT–6th Ed.	S 99	Raw	–	–	–	–	8	0.58*	–	–	8	0.85*	–	–
MAT–7th Ed.	S 99	SS	–	–	–	–	25	0.73*	17	0.76*	21	0.76	23	0.58*
Michigan Educational Assessment Program (MEAP) – English Language Arts														
MEAP	F 04	SS	–	–	–	–	–	–	155	0.81*	–	–	–	–
MEAP	F 05	SS	–	–	–	–	218	0.76*	196	0.80*	202	0.80*	207	0.69*
MEAP	F 06	SS	–	–	–	–	116	0.79*	132	0.69*	154	0.81*	129	0.66*
Michigan Educational Assessment Program (MEAP) – Reading														
MEAP	F 04	SS	–	–	–	–	–	–	155	0.80*	–	–	–	–
MEAP	F 05	SS	–	–	–	–	218	0.77*	196	0.78*	202	0.81*	207	0.68*
MEAP	F 06	SS	–	–	–	–	116	0.75*	132	0.70*	154	0.82*	129	0.70*
Mississippi Curriculum Test (MCT2)														
MCT2	S 02	SS	–	–	–	–	–	–	155	0.80*	–	–	–	–
MCT2	S 03	SS	–	–	–	–	218	0.77*	196	0.78*	202	0.81*	207	0.68*
MCT2	S 08	SS	–	–	–	–	3,821	0.74*	3,472	0.73*	2,915	0.71*	2,367	0.68*
Missouri Mastery Achievement Test (MMAT)														
MMAT	S 99	NCE	–	–	–	–	–	–	–	–	26	0.62*	–	–

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 64: Concurrent Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a]

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
New Jersey Assessment of Skills and Knowledge (NJ ASK)														
NJ ASK	S 13	SS	–	–	–	–	1,636	0.79*	1,739	0.80*	1,486	0.82*	440	0.77*
New York State Assessment Program														
NYSTP	S 13	SS	–	–	–	–	185	0.78*	–	–	–	–	–	–
North Carolina End-of-Grade (NCEOG): Test														
	S 99	SS	–	–	–	–	–	–	–	–	85	0.79*	–	–
NCEOG	S 06–08	SS	–	–	–	–	2,707	0.80*	2,234	0.77*	1,752	0.77*	702	0.77*
Ohio Achievement Assessment (OAA)														
OAA	S 13	SS	–	–	–	–	1,718	0.72*	1,595	0.71*	1,609	0.77*	1,599	0.76*
Oklahoma Core Curriculum Test (OCCT)														
OCCT	S 06	SS	–	–	–	–	78	0.62*	92	0.58*	46	0.52*	80	0.60*
OCCT	S 13	SS	–	–	–	–	153	0.79*	66	0.79*	72	0.80*	64	0.72*
South Dakota State Test of Educational Progress (DSTEP)														
DSTEP	S 08–10	SS	–	–	–	–	2,072	0.78*	1,751	0.77*	1,409	0.80*	906	0.78*
Stanford Achievement Test (SAT)														
SAT 9th Ed.	S 99	NCE	68	0.79*	–	–	26	0.44*	–	–	–	–	86	0.65*
SAT 9th Ed.	S 99	SS	11	0.89*	18	0.89*	67	0.79*	66	0.79*	72	0.80*	64	0.72*
State of Texas Assessments of Academic Readiness Standards Test (STAAR)														
STAAR	S 12–13	SS	–	–	–	–	8,567	0.79*	7,902	0.78*	7,272	0.76*	5,697	0.78*
Tennessee Comprehensive Assessment Program (TCAP)														
TCAP	S 11	SS	–	–	–	–	62	0.66*	56	0.59*	–	–	–	–
TCAP	S 12	SS	–	–	–	–	91	0.79*	118	0.21*	81	0.64*	–	–
TCAP	S 13	SS	–	–	–	–	494	0.73*	441	0.66*	426	0.77*	–	–
TerraNova														
TerraNova	S 99	SS	–	–	61	0.72*	117	0.78*	–	–	–	–	–	–
Texas Assessment of Academic Skills (TAAS)														
TAAS	S 99	NCE	–	–	–	–	–	–	–	–	–	–	229	0.66*
Transitional Colorado Assessment Program (TCAP)														
TCAP	S 12–13	SS	–	–	–	–	3,144	0.78*	3,200	0.82*	3,186	0.81*	3,106	0.83*

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 64: Concurrent Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a]

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
West Virginia Educational Standards Test 2 (WESTEST 2)														
WESTEST 2	S 12	SS	–	–	–	–	2,949	0.76*	7,537	0.77*	5,666	0.76*	2,390	0.75*
Woodcock Reading Mastery (WRM)														
	S 99		–	–	–	–	–	–	–	–	7	0.68*	7	0.66*
Wisconsin Knowledge and Concepts Examination (WKCE)														
WKCE	F 06–10	SS					8,649	0.78*	7,537	0.77*	5,666	0.76*	2,390	0.75*

Summary							
Grade(s)	**All**	**1**	**2**	**3**	**4**	**5**	**6**
Number of students	255,538	1,068	3,629	76,942	66,400	54,173	31,686
Number of coefficients	195	10	18	47	47	41	32
Average validity		0.80	0.73	0.72	0.72	0.74	0.72
Overall average	0.74						

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 65: Concurrent Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 7–12[a]

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Arkansas Augmented Benchmark Examination (AABE)														
AABE	S 08	SS	318	0.79*	278	0.76*	–	–	–	–	–	–	–	–
California Achievement Test (CAT)														
CAT/5	S 99	NCE	–	–	–	–	59	0.65*	–	–	–	–	–	–
CAT/5	S 99	SS	124	0.74*	131	0.76*	–	–	–	–	–	–	–	–
CAT/5	F 10–11	SS	146	0.75*	139	0.79*	92	0.64*	81	0.82*	48	0.79*	39	0.73*
Colorado Student Assessment Program (CSAP)														
CSAP	S 06	SS	299	0.84*	185	0.83*	–	–	–	–	–	–	–	–
Delaware Students Testing Program (DSTP) – Reading														
DSTP	S 05	SS	–	–	–	–	–	–	112	0.78*	–	–	–	–
DSTP	S 06	SS	150	0.72*	–	–	–	–	–	–	–	–	–	–

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 65: Concurrent Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 7–12[a]

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Florida Comprehensive Assessment Test (FCAT)														
FCAT	S 06	SS	–	–	74	0.65*	–	–	–	–	–	–	–	–
FCAT	S 06–08	SS	1,119	0.74*	618	0.76*	–	–	–	–	–	–	–	–
Florida Comprehensive Assessment Test (FCAT 2.0)														
FCAT 2.0	S 13	SS	158	0.83*	111	0.81*	–	–	–	–	–	–	–	–
Idaho Standards Achievement Test (ISAT)														
ISAT	S 06–08	SS	851	0.78*	895	0.71*	–	–	–	–	–	–	–	–
Illinois Standards Achievement Test (ISAT) – Reading														
ISAT	S 05	SS	–	–	157	0.73*	–	–	–	–	–	–	–	–
ISAT	S 06	SS	140	0.70*	–	–	–	–	–	–	–	–	–	–
Iowa Test of Basic Skills (ITBS)														
ITBS–K	S 99	NCE	–	–	–	–	67	0.78*	–	–	–	–	–	–
ITBS–L	S 99	SS	47	0.56*	–	–	65	0.64*	–	–	–	–	–	–
Kansas State Assessment Program (KSAP)														
KSAP	S 06–08	SS	1,147	0.70*	876	0.71*	–	–	–	–	–	–	–	–
Kentucky Core Content Test (KCCT)														
KCCT	S 08–10	SS	2,572	0.56*	1,198	0.56*	–	–	–	–	–	–	–	–
Michigan Educational Assessment Program – English Language Arts														
MEAP	F 04	SS	154	0.68*	–	–	–	–	–	–	–	–	–	–
MEAP	F 05	SS	233	0.72*	239	0.70*	–	–	–	–	–	–	–	–
MEAP	F 06	SS	125	0.79*	152	0.74*	–	–	–	–	–	–	–	–
Michigan Educational Assessment Program – Reading														
MEAP–R	F 04	SS	154	0.68*	–	–	–	–	–	–	–	–	–	–
MEAP–R	F 05	SS	233	0.72*	239	0.70*	–	–	–	–	–	–	–	–
MEAP–R	F 06	SS	125	0.79*	152	0.74*	–	–	–	–	–	–	–	–
Mississippi Curriculum Test (MCT2)														
MCT2	S 03	SS	372	0.70*	–	–	–	–	–	–	–	–	–	–
MCT2	S 08	SS	1,424	0.69*	1,108	0.72*	–	–	–	–	–	–	–	–
Missouri Mastery Achievement Test (MMAT)														
MMAT	S 99	NCE	–	–	29	0.78*	19	0.71*	–	–	–	–	–	–

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 65: Concurrent Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 7–12[a]

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
North Carolina End-of-Grade (NCEOG) Test														
NCEOG	S 06–08	SS	440	0.76*	493	0.74*	–	–	–	–	–	–	–	–
New Jersey Assessment of Skills and Knowledge (NJ ASK)														
NJ ASK	S 13	SS	595	0.78*	589	0.70*	–	–	–	–	–	–	–	–
Northwest Evaluation Association Levels Test (NWEA)														
NWEA-Achieve	S 99	NCE	–	–	124	0.66*	–	–	–	–	–	–	–	–
South Dakota State Test of Educational Progress (DSTEP)														
DSTEP	S 08–10	SS	917	0.78*	780	0.77*	–	–	–	–	–	–	–	–
Stanford Achievement Test (SAT)														
SAT–9th Ed.	S 99	NCE	50	0.65*	50	0.51*	–	–	–	–	–	–	–	–
SAT–9th Ed.	S 99	SS	70	0.70*	68	0.80*	–	–	–	–	–	–	–	–
State of Texas Assessments of Academic Readiness Standards Test (STAAR)														
STAAR	S 12–13	SS	5,062	.075*	4,651	0.75*	–	–	–	–	–	–	–	–
Test Achievement and Proficiency (TAP)														
TAP	S 99	NCE	–	–	–	–	6	0.42	13	0.80*	7	0.6	–	–
Texas Assessment of Academic Skills (TAAS)														
TAAS	S 99	NCE	–	–	–	–	–	–	43	0.60*	–	–	–	–
Transitional Colorado Assessment Program (TCAP)														
TCAP	S 12–13	SS	3,165	0.83*	3,106	0.83*	1,466	0.72*	–	–	–	–	–	–
West Virginia Educational Standards Test 2 (WESTEST 2)														
WESTEST 2	S 12	SS	1,612	0.76	1,396	0.75	–	–	–	–	–	–	–	–
Wisconsin Knowledge and Concepts Examination (WKCE)														
WKCE	F 06–10	SS	1,811	0.81	1,886	0.77	–	–	506	0.79	–	–	–	–
Wide Range Achievement Test 3 (WRAT3)														
WRAT3	S 99		–	–	17	0.81*	–	–	–	–	–	–	–	–

Summary

Grade(s)	All	7	8	9	10	11	12
Number of students	48,789	25,032	21,134	1,774	755	55	39
Number of coefficients	74	30	29	7	5	2	1
Average validity	–	0.74	0.73	0.65	0.76	0.70	0.73
Overall average	0.72						

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 66: Predictive Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 1–6[a]

Test Form	Date[b]	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
AIMSweb														
R-CBM	S 12	correct	60	0.14	156	0.38*	105	0.11	102	0.52*	–	–	–	–
Arkansas Augmented Benchmark Examination (AABE)														
AABE	F 07	SS	–	–	–	–	5,255	0.79*	5,208	0.77*	3,884	0.75*	3,312	0.75*
Colorado Student Assessment Program (CSAP)														
CSAP	F 04	–	–	–	–	–	82	0.72*	79	0.77*	93	0.70*	280	0.77*
Delaware Student Testing Program (DSTP) – Reading														
DSTP	S 05	–	–	–	–	–	189	0.58*	–	–	–	–	–	–
DSTP	W 05	–	–	–	–	–	120	0.67*	–	–	–	–	–	–
DSTP	S 05	–	–	–	–	–	161	0.52*	191	0.55*	190	0.62*	–	–
DSTP	F 05	–	–	–	253	0.64*	214	0.39*	256	0.62*	270	0.59*	242	0.71*
DSTP	W 05	–	–	–	275	0.61*	233	0.47*	276	0.59*	281	0.62*	146	0.57*
Florida Comprehensive Assessment Test (FCAT)														
FCAT	F 05	–	–	–	–	–	–	–	42	0.73*	–	–	409	0.67*
FCAT	W 07	–	–	–	–	–	–	–	–	–	–	–	417	0.76*
FCAT	F 05–07	SS	–	–	–	–	25,192	0.78*	21,650	0.75*	17,469	0.75*	9,998	0.73*
Florida Comprehensive Assessment Test (FCAT 2.0)														
FCAT 2.0	S 13	SS	–	–	–	–	6,788	0.78*	5,894	0.80*	5,374	0.80*	616	0.74*
Idaho Standards Achievement Test (ISAT)														
ISAT	F 08–10	SS	–	–	–	–	8,219	0.77*	8,274	0.77*	7,537	0.76*	5,742	0.77*
Illinois Standards Achievement Test (ISAT) – Reading														
ISAT–R	F 05	–	–	–	–	–	450	0.73*	–	–	317	0.68*	–	–
ISAT–R	W 05	–	–	–	–	–	564	0.76*	–	–	403	0.68*	–	–
ISAT–R	F 05	–	–	–	–	–	133	0.73*	140	0.74*	145	0.66*	–	–
ISAT–R	W 06	–	–	–	–	–	138	0.76*	145	0.77*	146	0.70*	–	–
Iowa Assessment														
IA	F 12	SS	–	–	–	–	1,763	0.61*	1,826	0.61*	1,926	0.59*	1,554	0.64*
IA	W 12	SS	–	–	–	–	548	0.60*	661	0.62*	493	0.64*	428	0.65*
IA	S 12	SS	–	–	–	–	1,808	0.63*	1,900	0.63*	1,842	0.65*	1,610	0.63*

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.
b. Dates correspond to the term and year of the predictor scores. With some exceptions, criterion scores were obtained during the same academic year. In some cases, data representing multiple years were combined. These dates are reported as a range (e.g. Fall 05–Fall 07).

Table 66: Predictive Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 1–6[a]

Test Form	Date[b]	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
Kentucky Core Content Test (KCCT)														
KCCT	F 07–09	SS	–	–	–	–	16,521	0.62*	15,143	0.57*	12,549	0.53*	9,091	0.58*
Michigan Educational Assessment Program (MEAP) – English Language Arts														
MEAP–EL	F 04	–	–	–	–	–	193	0.60*	181	0.70*	170	0.75*	192	0.66*
MEAP–EL	W 05	–	–	–	–	–	204	0.68*	184	0.74*	193	0.75*	200	0.70*
MEAP–EL	S 05	–	–	–	–	–	192	0.73*	171	0.73*	191	0.71*	193	0.62*
MEAP–EL	F 05	–	–	–	–	–	111	0.66*	132	0.71*	119	0.77*	108	0.60*
MEAP–EL	W 06	–	–	–	–	–	114	0.77*	–	–	121	0.75*	109	0.66*
Michigan Educational Assessment Program (MEAP) – Reading														
MEAP–R	F 04	–	–	–	–	–	193	0.60*	181	0.69*	170	0.76*	192	0.66*
MEAP–R	W 05	–	–	–	–	–	204	0.69*	184	0.74*	193	0.78*	200	0.70*
MEAP–R	S 05	–	–	–	–	–	192	0.72*	171	0.72*	191	0.74*	193	0.62*
MEAP–R	F 05	–	–	–	–	–	111	0.63*	132	0.70*	119	0.78*	108	0.62*
MEAP–R	W 06	–	–	–	–	–	114	0.72*	–	–	121	0.75*	109	0.64*
Mississippi Curriculum Test (MCT2)														
MCT2	F 01	–	–	–	86	0.57*	95	0.70*	97	0.65*	78	0.76*	–	–
MCT2	F 02	–	–	–	340	0.67*	337	0.67*	282	0.69*	407	0.71*	442	0.72*
MCT2	F 07	SS	–	–	–	–	6,184	0.77*	5,515	.74*	5,409	0.74*	4,426	0.68*
North Carolina End–of–Grade (NCEOG) Test														
NCEOG	F 05–07	SS	–	–	–	–	6,976	0.81*	6,531	0.78*	6,077	0.77*	3,255	0.77*
New York State Assessment Program														
NYSTP	S 13	SS	–	–	–	–	349	0.73*	–	–	–	–	–	–
Ohio Achievement Assessment														
OAA	S 13	SS	–	–	–	–	28	0.78*	41	0.52*	29	0.79*	30	0.75*
Oklahoma Core Curriculum Test (OCCT)														
OCCT	F 04	–	–	–	–	–	–	–	–	–	44	0.63*	–	–
OCCT	W 05	–	–	–	–	–	–	–	–	–	45	0.66*	–	–
OCCT	F 05	–	–	–	–	–	89	0.59*	90	0.60*	79	0.69*	84	0.63*
OCCT	W 06	–	–	–	–	–	60	0.65*	40	0.67*	–	–	–	–

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.
b. Dates correspond to the term and year of the predictor scores. With some exceptions, criterion scores were obtained during the same academic year. In some cases, data representing multiple years were combined. These dates are reported as a range (e.g. Fall 05–Fall 07).

Table 66: Predictive Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 1–6[a]

Test Form	Date[b]	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
South Dakota State Test of Educational Progress (DSTEP)														
DSTEP	F 07–09	SS	–	–	–	–	3,909	0.79*	3,679	0.78*	3,293	0.78*	2,797	0.79*
Star Reading														
Star–R	F 05	–	16,982	0.66*	42,601	0.78*	46,237	0.81*	44,125	0.83*	34,380	0.83*	23,378	0.84*
Star–R	F 06	–	25,513	0.67*	63,835	0.78*	69,835	0.81*	65,157	0.82*	57,079	0.83*	35,103	0.83*
Star–R	F 05	–	8,098	0.65*	20,261	0.79*	20,091	0.81*	18,318	0.82*	7,621	0.82*	5,021	0.82*
Star–R	F 05	–	8,098	0.55*	20,261	0.72*	20,091	0.77*	18,318	0.80*	7,621	0.80*	5,021	0.79*
Star–R	S 06	–	8,098	0.84*	20,261	0.82*	20,091	0.83*	18,318	0.83*	7,621	0.83*	5,021	0.83*
Star–R	S 06	–	8,098	0.79*	20,261	0.80*	20,091	0.81*	18,318	0.82*	7,621	0.82*	5,021	0.81*
State of Texas Assessments of Academic Readiness Standards Test (STAAR)														
STAAR	S 12–13	SS	–	–	–	–	6,132	0.81*	5,744	0.80*	5,327	0.79*	5,143	0.79*
Tennessee Comprehensive Assessment Program (TCAP)														
TCAP	S 11	SS	–	–	–	–	695	0.68*	602	0.72*	315	0.61*	–	–
TCAP	S 12	SS	–	–	–	–	763	0.70*	831	0.33*	698	0.65*	–	–
TCAP	S 13	SS	–	–	–	–	2,509	0.67*	1,897	0.63*	1,939	0.68*	431	0.65*
West Virginia Educational Standards Test 2 (WESTEST 2)														
WESTEST 2	S 12	SS	–	–	–	–	2,828	0.80*	3,078	0.73*	3,246	0.73*	3,214	0.73*
Wisconsin Knowledge and Concepts Examination (WKCE)														
WKCE	S 05–09	SS					15,706	0.75*	15,569	0.77*	13,980	0.78*	10,641	0.78*

Summary							
Grade(s)	**All**	**1**	**2**	**3**	**4**	**5**	**6**
Number of students	1,227,887	74,887	188,434	313,102	289,571	217,416	144,477
Number of coefficients	194	6	10	49	43	47	39
Average validity		0.69	0.72	0.70	0.71	0.72	0.71
Overall average	0.71						

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.
b. Dates correspond to the term and year of the predictor scores. With some exceptions, criterion scores were obtained during the same academic year. In some cases, data representing multiple years were combined. These dates are reported as a range (e.g. Fall 05–Fall 07).

Table 67: Predictive Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 7–12[a]

Test Form	Date[b]	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Arkansas Augmented Benchmark Examination (AABE)														
AABE	F 07	SS	2,418	0.74*	1,591	0.75*	–	–	–	–	–	–	–	–
Colorado Student Assessment Program (CSAP)														
CSAP	F 05	–	299	0.83*	185	0.83*	–	–	–	–	–	–	–	–
Delaware Student Testing Program (DSTP) – Reading														
DSTP	S 05	–	100	0.75*	143	0.63*	–	–	48	0.66*	–	–	–	–
DSTP	F 05	–	273	0.69*	247	0.70*	152	0.73*	97	0.78*	–	–	–	–
DSTP	W 05	–	–	–	61	0.64*	230	0.64*	145	0.71*	–	–	–	–
Florida Comprehensive Assessment Test (FCAT)														
FCAT	F 05	–	381	0.61*	387	0.62*	–	–	–	–	–	–	–	–
FCAT	W 07	–	342	0.64*	361	0.72*	–	–	–	–	–	–	–	–
FCAT	F 05–07	SS	8,525	0.72*	6,216	0.72*	–	–	–	–	–	–	–	–
Florida Comprehensive Assessment Test (FCAT 2.0)														
FCAT 2.0	S 13	SS	586	0.75*	653	0.78*	–	–	–	–	–	–	–	–
Idaho Standards Achievement Test (ISAT)														
ISAT	F 05–07	SS	4,119	0.76*	3,261	0.73*	–	–	–	–	–	–	–	–
Illinois Standards Achievement Test (ISAT) – Reading														
ISAT	F 05	–	173	0.51*	158	0.66*	–	–	–	–	–	–	–	–
Iowa Assessment														
IA	F 12	SS	1,264	0.60*	905	0.63*	–	–	–	–	–	–	–	–
IA	W 12	SS	118	0.66*	72	0.67*	–	–	–	–	–	–	–	–
IA	S 12	SS	1,326	0.68*	1,250	0.66*	–	–	–	–	–	–	–	–
Kentucky Core Content Test (KCCT)														
KCCT	F 07–09	SS	4,962	0.57*	2,530	0.58*	–	–	–	–	–	–	–	–

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.
b. Dates correspond to the term and year of the predictor scores. With some exceptions, criterion scores were obtained during the same academic year. In some cases, data representing multiple years were combined. These dates are reported as a range (e.g. Fall 05–Fall 07).

Table 67: Predictive Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 7–12[a]

Test Form	Date[b]	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Michigan Educational Assessment Program (MEAP) – English Language Arts														
MEAP	F 04	–	181	0.71*	88	0.85*	–	–	–	–	–	–	–	–
MEAP	W 05	–	214	0.73*	212	0.73*	–	–	–	–	–	–	–	–
MEAP	S 05	–	206	0.75*	223	0.69*	–	–	–	–	–	–	–	–
MEAP	F 05	–	114	0.66*	126	0.66*	–	–	–	–	–	–	–	–
MEAP	W 06	–	114	0.64*	136	0.71*	–	–	–	–	–	–	–	–
MEAP	S 06	–	–	–	30	0.80*	–	–	–	–	–	–	–	–
Michigan Educational Assessment Program (MEAP) – Reading														
MEAP–R	F 04	–	181	0.70*	88	0.84*	–	–	–	–	–	–	–	–
MEAP–R	W 05	–	214	0.72*	212	0.73*	–	–	–	–	–	–	–	–
MEAP–R	S 05	–	206	0.72*	223	0.69*	–	–	–	–	–	–	–	–
MEAP–R	F 05	–	116	0.68*	138	0.66*	–	–	–	–	–	–	–	–
MEAP–R	W 06	–	116	0.68*	138	0.70*	–	–	–	–	–	–	–	–
MEAP–R	S 06	–	–	–	30	0.81*	–	–	–	–	–	–	–	–
Mississippi Curriculum Test (MCT2)														
MCT2	F 02	–	425	0.68*	–	–	–	–	–	–	–	–	–	–
MCT2	F 07	SS	3,704	0.68*	3,491	0.73*	–	–	–	–	–	–	–	–
North Carolina End–of–Grade (NCEOG) Test														
NCEOG	F 05–07	SS	2,735	0.77*	2,817	0.77*	–	–	–	–	–	–	–	–
Ohio Achievement Assessment														
OAA	S 13	SS	53	0.82*	38	0.66*	–	–	–	–	–	–	–	–
South Dakota State Test of Educational Progress (DSTEP)														
DSTEP	F 07–09	SS	2,236	0.79*	2,073	0.78*	–	–	–	–	–	–	–	–
Star Reading														
Star–R	F 05	–	17,370	0.82*	9,862	0.82*	2,462	0.82*	15,277	0.85*	1,443	0.83*	596	0.85*
Star–R	F 06	–	22,177	0.82*	19,152	0.82*	4,087	0.84*	2,624	0.85*	2,930	0.85*	2,511	0.86*
Star–R	F 05	–	5,399	0.81*	641	0.76*	659	0.89*	645	0.88*	570	0.90*	–	–
Star–R	F 05	–	5,399	0.79*	641	0.76*	659	0.83*	645	0.83*	570	0.87*	–	–
Star–R	S 06	–	5,399	0.82*	641	0.83*	659	0.87*	645	0.88*	570	0.89*	–	–
Star–R	S 06	–	5,399	0.80*	641	0.83*	659	0.85*	645	0.85*	570	0.86*		

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.
b. Dates correspond to the term and year of the predictor scores. With some exceptions, criterion scores were obtained during the same academic year. In some cases, data representing multiple years were combined. These dates are reported as a range (e.g. Fall 05–Fall 07).

Table 67: Predictive Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 7–12[a]

Test Form	Date[b]	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
State of Texas Assessments of Academic Readiness Standards Test (STAAR)														
STAAR	S 12–13	SS	4,716	0.77*	4,507	0.76*	–	–	–	–	–	–	–	–
Tennessee Comprehensive Assessment Program (TCAP)														
TCAP	S 13	SS	332	0.81*	233	0.74*	–	–	–	–	–	–	–	–
West Virginia Educational Standards Test 2 (WESTEST 2)														
WESTEST 2	S 12	SS	2,852	0.71*	2,636	0.74*	–	–	–	–	–	–	–	–
Wisconsin Knowledge and Concepts Examination (WKCE)														
WKCE	S 05–09	SS	6,399	0.78*	5,500	0.78*			401	0.78*				

Summary

Grade(s)	All	7	8	9	10	11	12
Number of students	224,179	111,143	72,537	9,567	21,172	6,653	3,107
Number of coefficients	106	39	41	8	10	6	2
Average validity	–	0.72	0.73	0.81	0.81	0.87	0.86
Overall average	0.80						

a. * Denotes correlation coefficients that are statistically significant at the 0.05 level.
b. Dates correspond to the term and year of the predictor scores. With some exceptions, criterion scores were obtained during the same academic year. In some cases, data representing multiple years were combined. These dates are reported as a range (e.g. Fall 05–Fall 07).

Table 68: Other External Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6[a]

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
American Testronics														
Level C-3	Spr 98	Scaled	–	–	20	0.71*	–	–	–	–	–	–	–	–
California Achievement Test (CAT)														
/ 4	Spr 98	Scaled	–	–	16	0.82*	–	–	54	0.65*	–	–	10	0.88*
/ 5	Spr 98	Scaled	–	–	–	–	40	0.82*	103	0.85*	–	–	–	–
/ 5	Fall 98	NCE	40	0.83*	–	–	–	–	–	–	–	–	–	–
/ 5	Fall 98	Scaled	–	–	–	–	39	0.85*	–	–	–	–	–	–
Comprehensive Test of Basic Skills (CTBS)														
A-15	Fall 97	NCE	–	–	–	–	–	–	–	–	–	–	24	0.79*
/ 4	Spr 97	Scaled	–	–	–	–	–	–	–	–	31	0.61*	–	–
/ 4	Spr 98	Scaled	–	–	–	–	–	–	6	0.49	68	0.76*	–	–
A-19/20	Spr 98	Scaled	–	–	–	–	–	–	–	–	10	0.73*	–	–
A-15	Spr 98	Scaled	–	–	–	–	–	–	–	–	–	–	93	0.81*
A-16	Fall 98	NCE	–	–	–	–	–	–	–	–	–	–	73	0.67*
Degrees of Reading Power (DRP)														
	Spr 98		–	–	–	–	8	0.71*	–	–	25	0.72*	23	0.38
Gates-MacGinitie Reading Test (GMRT)														
2nd Ed., D	Spr 98	NCE	–	–	–	–	–	–	–	–	–	–	47	0.80*
L-3rd	Spr 98	NCE	–	–	31	0.69*	27	0.62*	–	–	–	–	–	–
L-3rd	Fall 98	NCE	60	0.64*	–	–	66	0.83*	–	–	–	–	–	–
Indiana Statewide Testing for Educational Progress (ISTEP)														
	Fall 98	NCE	–	–	–	–	19	0.80*	–	–	–	–	21	0.79*
Iowa Test of Basic Skills (ITBS)														
Form K	Spr 98	NCE	–	–	–	–	88	0.74*	17	0.59*	–	–	21	0.83*
Form L	Spr 98	NCE	–	–	–	–	50	0.84*	–	–	–	–	57	0.66*
Form M	Spr 98	NCE	–	–	68	0.71*	–	–	–	–	–	–	–	–
Form K	Fall 98	NCE	–	–	67	0.66*	43	0.73*	67	0.74*	28	0.81*	–	–
Form L	Fall 98	NCE	–	–	–	–	–	–	27	0.88*	6	0.97*	37	0.60*
Form M	Fall 98	NCE	–	–	65	0.81*	–	–	53	0.72*	–	–	–	–

a. Sample sizes are in the columns labeled "n."
* Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 68: Other External Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6[a]

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
Metropolitan Achievement Test (MAT)														
7th Ed.	Spr 98	NCE	–	–	–	–	–	–	29	0.67*	22	0.68*	17	0.86*
6th Ed	Spr 98	Raw	–	–	–	–	–	–	6	0.91*	–	–	5	0.67
7th Ed.	Spr 98	Scaled	–	–	48	0.75*	–	–	–	–	30	0.79*	–	–
7th Ed.	Fall 98	NCE	–	–	–	–	–	–	–	–	–	–	49	0.75*
Metropolitan Readiness Test (MRT)														
	Spr 96	NCE	–	–	–	–	5	0.81	–	–	–	–	–	–
	Spr 98	NCE	4	0.63	–	–	–	–	–	–	–	–	–	–
Missouri Mastery Achievement Test (MMAT)														
	Spr 98	Scaled	–	–	–	–	12	0.44	–	–	14	0.75*	24	0.62*
New York State Pupil Evaluation Program (P&P)														
	Spr 98		–	–	–	–	–	–	13	0.92*	–	–	–	–
North Carolina End of Grade Test (NCEOG)														
	Spr 98	Scaled	–	–	–	–	–	–	–	–	53	0.76*	–	–
NRT Practice Achievement Test (NRT)														
Practice	Spr 98	NCE	–	–	56	0.71*	–	–	–	–	–	–	–	–
Stanford Achievement Test (Stanford)														
9th Ed.	Spr 97	Scaled	–	–	–	–	–	–	–	–	68	0.65*	–	–
7th Ed.	Spr 98	Scaled	11	0.73*	7	0.94*	8	0.65	15	0.82*	7	0.87*	8	0.87*
8th Ed.	Spr 98	Scaled	8	0.94*	8	0.64	6	0.68	11	0.76*	8	0.49	7	0.36
9th Ed.	Spr 98	Scaled	13	0.73*	93	0.73*	19	0.62*	314	0.74*	128	0.72*	62	0.67*
4th Ed. 3/V	Spr 98	Scaled	14	0.76*	–	–	–	–	–	–	–	–	–	–
9th Ed.	Fall 98	NCE	–	–	–	–	45	0.89*	–	–	35	0.68*	–	–
9th Ed.	Fall 98	Scaled	–	–	88	0.60*	25	0.79*	–	–	196	0.73*	–	–
9th Ed. 2/ SA	Fall 98	Scaled	–	–	–	–	103	0.69*	–	–	–	–	–	–
Tennessee Comprehensive Assessment Program (TCAP)														
	Spr 98	Scaled	–	–	30	0.75*	–	–	–	–	–	–	–	–

a. Sample sizes are in the columns labeled "n."
* Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 68: Other External Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6[a]

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
TerraNova														
	Fall 97	Scaled	–	–	–	–	–	–	–	–	56	0.70*	–	–
	Spr 98	NCE	–	–	–	–	76	0.63*	–	–	–	–	–	–
	Spr 98	Scaled	–	–	94	0.50*	55	0.79*	299	0.75*	86	0.75*	23	0.59*
	Fall 98	NCE	–	–	–	–	–	–	–	–	–	–	126	0.74*
	Fall 98	Scaled	–	–	–	–	–	–	14	0.70*	–	–	15	0.77*
Wide Range Achievement Test 3 (WRAT3)														
	Fall 98		–	–	–	–	–	–	–	–	–	–	10	0.89*
Wisconsin Reading Comprehension Test														
	Spr 98		–	–	–	–	–	–	63	0.58*	–	–	–	–

Summary							
Grade(s)	**All**	**1**	**2**	**3**	**4**	**5**	**6**
Number of students	4,289	150	691	734	1,091	871	752
Number of coefficients	95	7	14	19	16	18	21
Average validity	–	0.75	0.72	0.73	0.74	0.73	0.71
Overall average	0.73						

a. Sample sizes are in the columns labeled "n."
* Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 69: Other External Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 7–12[a]

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
California Achievement Test (CAT)														
/ 4	Spr 98	Scaled	–	–	11	0.75*	–	–	–	–	–	–	–	–
/ 5	Spr 98	NCE	80	0.85*	–	–	–	–	–	–	–	–	–	–
Comprehensive Test of Basic Skills (CTBS)														
/ 4	Spr 97	NCE	–	–	12	0.68*	–	–	–	–	–	–	–	–
/ 4	Spr 98	NCE	43	0.84*	–	–	–	–	–	–	–	–	–	–
/ 4	Spr 98	Scaled	107	0.44*	15	0.57*	43	0.86*	–	–	–	–	–	–
A-16	Spr 98	Scaled	24	0.82*	–	–	–	–	–	–	–	–	–	–

a. Sample sizes are in the columns labeled "n."
* Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 69: Other External Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 7–12[a]

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Explore (ACT Program for Educational Planning, 8th Grade)														
	Fall 97	NCE	–	–	–	–	67	0.72*	–	–	–	–	–	–
	Fall 98	NCE	–	–	32	0.66*	–	–	–	–	–	–	–	–
Iowa Test of Basic Skills (ITBS)														
Form K	Spr 98	NCE	–	–	–	–	35	0.84*	–	–	–	–	–	–
Form K	Fall 98	NCE	32	0.87*	43	0.61*	–	–	–	–	–	–	–	–
Form K	Fall 98	Scaled	72	0.77*	67	0.65*	77	0.78*	–	–	–	–	–	–
Form L	Fall 98	NCE	19	0.78*	13	0.73*	–	–	–	–	–	–	–	–
Metropolitan Achievement Test (MAT)														
7th Ed.	Spr 97	Scaled	114	0.70*	–	–	–	–	–	–	–	–	–	–
7th Ed.	Spr 98	NCE	46	0.84*	63	0.86*	–	–	–	–	–	–	–	–
7th Ed.	Spr 98	Scaled	88	0.70*	–	–	–	–	–	–	–	–	–	–
7th Ed.	Fall 98	NCE	50	0.55*	48	0.75*	–	–	–	–	–	–	–	–
Missouri Mastery Achievement Test (MMAT)														
	Spr 98	Scaled	24	0.62*	12	0.72*	–	–	–	–	–	–	–	–
North Carolina End of Grade Test (NCEOG)														
	Spr 97	Scaled	–	–	–	–	–	–	58	0.81*	–	–	–	–
	Spr 98	Scaled	–	–	–	–	73	0.57*	–	–	–	–	–	–
PLAN (ACT Program for Educational Planning, 10th Grade)														
	Fall 97	NCE	–	–	–	–	–	–	–	–	46	0.71*	–	–
	Fall 98	NCE	–	–	–	–	–	–	104	0.53*	–	–	–	–
Preliminary Scholastic Aptitude Test (PSAT)														
	Fall 98	Scaled	–	–	–	–	–	–	–	–	78	0.67*	–	–
Stanford Achievement Test (Stanford)														
9th Ed.	Spr 97	Scaled	–	–	–	–	–	–	–	–	–	–	11	0.90*
7th Ed.	Spr 98	Scaled	–	–	8	0.83*	–	–	–	–	–	–	–	–
8th Ed.	Spr 98	Scaled	6	0.89*	8	0.78*	91	0.62*	–	–	93	0.72*	–	–
9th Ed.	Spr 98	Scaled	72	0.73*	78	0.71*	233	0.76*	32	0.25	64	0.76*	–	–
4th Ed. 3/V	Spr 98	Scaled	–	–	–	–	–	–	55	0.68*	–	–	–	–
9th Ed.	Fall 98	NCE	92	0.67*	–	–	–	–	–	–	–	–	–	–
9th Ed.	Fall 98	Scaled	–	–	–	–	93	0.75*	–	–	–	–	70	0.75*

a. Sample sizes are in the columns labeled "n."
* Denotes correlation coefficients that are statistically significant at the 0.05 level.

Table 69: Other External Validity Data: Star Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 7–12[a]

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Stanford Reading Test														
3rd Ed.	Fall 97	NCE	–	–	–	–	5	0.81	24	0.82*	–	–	–	–
TerraNova														
	Fall 97	NCE	103	0.69*	–	–	–	–	–	–	–	–	–	–
	Spr 98	Scaled	–	–	87	0.82*	–	–	21	0.47*	–	–	–	–
	Fall 98	NCE	35	0.69*	32	0.74*	–	–	–	–	–	–	–	–
Test of Achievement and Proficiency (TAP)														
	Spr 97	NCE	–	–	–	–	–	–	–	–	36	0.59*	–	–
	Spr 98	NCE	–	–	–	–	–	–	41	0.66*	–	–	43	0.83*
Texas Assessment of Academic Skills (TAAS)														
	Spr 97	TLI	–	–	–	–	–	–	–	–	–	–	41	0.58*
Wide Range Achievement Test 3 (WRAT3)														
	Spr 98		9	0.35	–	–	–	–	–	–	–	–	–	–
	Fall 98		–	–	–	–	16	0.80*	–	–	–	–	–	–
Wisconsin Reading Comprehension Test														
	Spr 98		–	–	–	–	–	–	63	0.58*	–	–	–	–

Summary

Grade(s)	All	7	8	9	10	11	12
Number of students	3,158	1,016	529	733	398	317	165
Number of coefficients	60	18	15	10	8	5	4
Average validity	–	0.71	0.72	0.75	0.60	0.69	0.77
Overall average	0.71						

a. Sample sizes are in the columns labeled "n."
* Denotes correlation coefficients that are statistically significant at the 0.05 level.

Relationship of Star Reading Scores to Scores on State Tests of Accountability in Reading

In the US, following the passage of the No Child Left Behind Act (NCLB) in 2001, all states moved to comprehensive tests of grade level standards for purposes of accountability. This created interest in the degree to which Star Reading test scores are related to state accountability test scores. The following section provides specific information about the validity of Star scores relative to state test scores of the NCLB era. Results of concurrent and predictive validity (defined earlier) are presented here with specific results for a variety of state tests of accountability.

Table 70 and Table 71 provide a variety of concurrent and predictive validity coefficients, respectively, for grades 3–8. Numerous state accountability tests have been used in this research.

Table 70: Concurrent Validity Data: Star Reading 2 Correlations (r) with State Accountability Tests, Grades 3–8[a]

Date	Score	3		4		5		6		7		8	
		n	r	n	r	n	r	n	r	n	r	n	r
Colorado Student Assessment Program													
Spr 06	Scaled	82	0.75*	79	0.83*	93	0.68*	280	0.80*	299	0.84*	185	0.83*
Delaware Student Testing Program—Reading													
Spr 05	Scaled	104	0.57*	–	–	–	–	–	–	–	–	–	–
Spr 06	Scaled	126	0.43*	141	0.62*	157	0.59*	75	0.66*	150	0.72	–	–
Florida Comprehensive Assessment Test													
Spr 06	SSS	–	–	41	0.65*	–	–	–	–	–	–	74	0.65*
Illinois Standards Achievement Test—Reading													
Spr 05	Scaled	594	0.76*	–	–	449	0.70*	–	–	–	–	157	0.73*
Spr 06	Scaled	140	0.80*	144	0.80*	146	0.72*	–	–	140	0.70*	–	–
Michigan Educational Assessment Program—English Language Arts													
Fall 04	Scaled	–	–	155	0.81*	–	–	–	–	154	0.68*	–	–
Fall 05	Scaled	218	0.76*	196	0.80*	202	0.80*	207	0.69*	233	0.72*	239	0.70*
Fall 06	Scaled	116	0.79*	132	0.69*	154	0.81*	129	0.66*	125	0.79*	152	0.74*

a. Sample sizes are in the columns labeled "n."
* Denotes correlation coefficients that are statistically significant ($p < 0.05$).

Table 70: Concurrent Validity Data: Star Reading 2 Correlations (r) with State Accountability Tests, Grades 3–8[a]

Date	Score	3		4		5		6		7		8	
		n	r	n	r	n	r	n	r	n	r	n	r
Michigan Educational Assessment Program—Reading													
Fall 04	Scaled	–	–	155	0.80*	–	–	–	–	156	0.68*	–	–
Fall 05	Scaled	218	0.77*	196	0.78*	202	0.81*	207	0.68*	233	0.71*	239	0.69*
Fall 06	Scaled	116	0.75*	132	0.70*	154	0.82*	129	0.70*	125	0.86*	154	0.72*
Mississippi Curriculum Test													
Spr 02	Scaled	148	0.62*	175	0.66*	81	0.69*	–	–	–	–	–	–
Spr 03	Scaled	389	0.71*	359	0.70*	377	0.70*	364	0.72*	372	0.70*	–	–
Oklahoma Core Curriculum Test													
Spr 06	Scaled	78	0.62*	92	0.58*	46	0.52*	80	0.60*	–	–	–	–

Summary							
Grades	**All**	**3**	**4**	**5**	**6**	**7**	**8**
Number of students	11,045	2,329	1,997	2,061	1,471	1,987	1,200
Number of coefficients	61	12	13	11	8	10	7
Average validity	–	0.72	0.73	0.73	0.71	0.74	0.73
Overall validity	0.73						

a. Sample sizes are in the columns labeled "n."
* Denotes correlation coefficients that are statistically significant ($p < 0.05$).

Table 71: Predictive Validity Data: Star Reading Scaled Scores Predicting Later Performance for Grades 3–8 on Numerous State Accountability Tests[a]

Predictor Date	Criterion Date[b]	3		4		5		6		7		8	
		n	r	n	r	n	r	n	r	n	r	n	r
Colorado Student Assessment Program													
Fall 05	Spr 06	82	0.72*	79	0.77*	93	0.70*	280	0.77*	299	0.83*	185	0.83*
Delaware Student Testing Program—Reading													
Fall 04	Spr 05	189	0.58*	–	–	–	–	–	–	–	–	–	–
Win 05	Spr 05	120	0.67*	–	–	–	–	–	–	–	–	–	–
Spr 05	Spr 06	161	0.52*	191	0.55*	190	0.62*	–	–	100	0.75*	143	0.63*
Fall 05	Spr 06	214	0.39*	256	0.62*	270	0.59*	242	0.71*	273	0.69*	247	0.70*
Win 05	Spr 06	233	0.47*	276	0.59*	281	0.62*	146	0.57*	–	–	61	0.64*
Florida Comprehensive Assessment Test													
Fall 05	Spr 06	–	–	42	0.73*	–	–	409	0.67*	381	0.61*	387	0.62*
Win 07	Spr 07	–	–	–	–	–	–	417	0.76*	342	0.64*	361	0.72*
Illinois Standards Achievement Test—Reading													
Fall 04	Spr 05	450	0.73*	–	–	317	0.68*	–	–	–	–	–	–
Win 05	Spr 05	564	0.76*	–	–	403	0.68*	–	–	–	–	–	–
Fall 05	Spr 06	133	0.73*	140	0.74*	145	0.66*	–	–	173	0.51*	158	0.66*
Win 06	Spr 06	138	0.76*	145	0.77*	146	0.70*	–	–	–	–	–	–
Michigan Educational Assessment Program—English Language Arts													
Fall 04	Fall 05[P]	193	0.60*	181	0.70*	170	0.75*	192	0.66*	181	0.71*	88	0.85*
Win 05	Fall 05[P]	204	0.68*	184	0.74*	193	0.75*	200	0.70*	214	0.73*	212	0.73*
Spr 05	Fall 05[P]	192	0.73*	171	0.73*	191	0.71*	193	0.62*	206	0.75*	223	0.69*
Fall 05	Fall 06[P]	111	0.66*	132	0.71*	119	0.77*	108	0.60*	114	0.66*	126	0.66*
Win 06	Fall 06[P]	114	0.77*	–	–	121	0.75*	109	0.66*	114	0.64*	136	0.71*
Spr 06	Fall 06[P]	–	–	–	–	–	–	–	–	–	–	30	0.80*
Michigan Educational Assessment Program—Reading													
Fall 04	Fall 05[P]	193	0.60*	181	0.69*	170	0.76*	192	0.66*	181	0.70*	88	0.84*
Win 05	Fall 05[P]	204	0.69*	184	0.74*	193	0.78*	200	0.70*	214	0.72*	212	0.73*
Spr 05	Fall 05[P]	192	0.72*	171	0.72*	191	0.74*	193	0.62*	206	0.72*	223	0.69*
Fall 05	Fall 06[P]	111	0.63*	132	0.70*	119	0.78*	108	0.62*	116	0.68*	138	0.66*
Win 06	Fall 06[P]	114	0.72*	–	–	121	0.75*	109	0.64*	116	0.68*	138	0.70*
Spr 06	Fall 06[P]	–	–	–	–	–	–	–	–	–	–	30	0.81*

a. Grade given in the column signifies the grade within which the Predictor variable was given (as some validity estimates span contiguous grades).
b. [P] indicates a criterion measure was given in a subsequent grade from the predictor.
* Denotes significant correlation ($p < 0.05$).

Table 71: Predictive Validity Data: Star Reading Scaled Scores Predicting Later Performance for Grades 3–8 on Numerous State Accountability Tests[a]

Predictor Date	Criterion Date[b]	3		4		5		6		7		8	
		n	r	n	r	n	r	n	r	n	r	n	r
Mississippi Curriculum Test													
Fall 01	Spr 02	95	0.70*	97	0.65*	78	0.76*	–	–	–	–	–	–
Fall 02	Spr 03	337	0.67*	282	0.69*	407	0.71*	442	0.72*	425	0.68*	–	–
Oklahoma Core Curriculum Test													
Fall 04	Spr 05	–	–	–	–	44	0.63*	–	–	–	–	–	–
Win 05	Spr 05	–	–	–	–	45	0.66*	–	–	–	–	–	–
Fall 05	Spr 06	89	0.59*	90	0.60*	79	0.69*	84	0.63*	–	–	–	–
Win 06	Spr 06	60	0.65*	40	0.67*	–	–	–	–	–	–	–	–

Summary

Grades	All	3	4	5	6	7	8
Number of students	22,018	4,493	2,974	4,086	3,624	3,655	3,186
Number of coefficients	119	24	19	23	17	17	19
Average validity	–	0.66	0.68	0.70	0.68	0.69	0.70
Overall validity	0.68						

a. Grade given in the column signifies the grade within which the Predictor variable was given (as some validity estimates span contiguous grades).
b. P indicates a criterion measure was given in a subsequent grade from the predictor.
* Denotes significant correlation ($p < 0.05$).

Relationship of Star Reading Enterprise Scores to Scores on Previous Versions

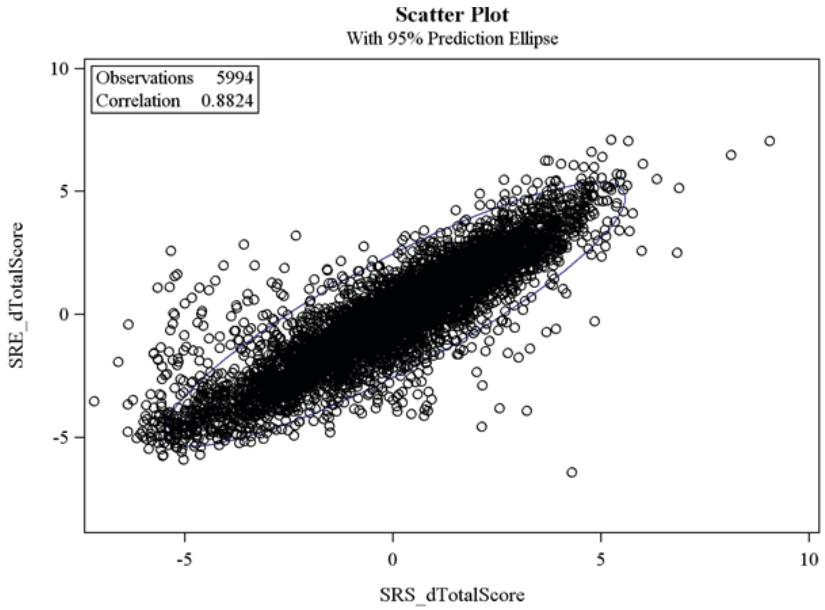
The 34-item version of Star Reading represents a significant departure from previous versions of Star. It is not a replacement for earlier versions; instead, it presents an alternative approach to reading assessment. Unlike previous Star Reading versions, which were primarily designed as measures only of reading comprehension, the 34-item version of Star Reading, simply referred to as Star Reading, is a standards-based assessment which measures a wide variety of reading skills. In addition to this substantial change in content from previous versions, Star Reading tests are also longer, and as a result have greater measurement precision and reliability.

Star Reading was released for use in June 2011. In the course of its development, Star Reading was administered to thousands of students who also took previous versions. The correlations between Star Reading and previous versions of Star Reading provide validity evidence of their own. To the extent that those correlations are high, they would provide evidence that the current Star Reading and previous versions are measuring the same or highly similar underlying attributes, even though they are dissimilar in content and measurement precision. Table 72 displays data on the correlations between Star Reading and scores on two previous versions: classic versions of Star Reading (which includes versions 2.0 through 4.3) and Star Reading Progress Monitoring (version 4.4.) Both of those Star Reading versions are 25-item versions that are highly similar to one another, differing primarily in terms of the software that delivers them; for all practical purposes, they may be considered alternate forms of Star Reading.

Table 72: Correlations of Star Reading with Scores on Star Reading Classic and Star Reading Progress Monitoring Tests

	Star Reading Classic Versions		Star Reading Progress Monitoring Version	
Grade	**N**	**r**	**N**	**r**
1	810	0.73	539	0.87
2	1,762	0.81	910	0.85
3	2,830	0.81	1,140	0.83
4	2,681	0.81	1,175	0.82
5	2,326	0.80	919	0.82
6	1,341	0.85	704	0.84
7	933	0.76	349	0.81
8	811	0.80	156	0.85
9	141	0.76	27	0.75
10	107	0.79	20	0.84
11	84	0.87	6	0.94
12	74	0.78	5	0.64
All Grades Combined	13,979	0.87	5,994	0.88

Figure 5: Scatterplot of Star Reading and Star Reading Progress Monitoring Test Scores for 5,994 Students Tested in June and July 2011



Data from Post-Publication Studies

Subsequent to publication of Star Reading 2.0 in 1999, additional external validity data became available, both from users of the assessment and from special studies conducted by Renaissance Learning and others. This section provides a summary of results of a doctoral dissertation examining the relationship of Star Reading to scores on a leading nationally standardized reading assessment, the Stanford Achievement Test (SAT9), and a major reading state test, the California Standards Test (CST).

Predictive Validity: Correlations with SAT9 and the California Standards Tests

A doctoral dissertation (Bennicoff-Nan, 2002) studied the validity of Star Reading as a predictor of student's scores in a California school district on the California Standards Test (CST) and the Stanford Achievement Tests, Ninth Edition (SAT9), the reading accountability tests mandated by the State of California. At the time of the study, those two tests were components of the California Standardized Testing and Reporting Program. The study involved analysis of test scores of more than 1,000 school children in four grades in a rural central California school district; 83 percent of students in the district were eligible for free and reduced lunch and 30 percent were identified as having limited English proficiency.

Bennicoff-Nan's dissertation addressed a number of different research questions. For purposes of this technical manual, we are primarily interested in the correlations between Star Reading 2 with SAT9 and CST scores. Those correlations are displayed by grade in Table 73.

Table 73: Correlations of Star Reading 2.0 Scores with SAT9 and California Standards Test Scores, by Grade

Grade	SAT9 Total Reading	CST English and Language Arts
3	0.82	0.78
4	0.83	0.81
5	0.83	0.79
6	0.81	0.78

In summary, the average correlation between Star Reading and SAT9 was 0.82. The average correlation with CST was 0.80. These values are evidence of the validity of Star Reading for predicting performance on both norm-referenced reading tests such as the SAT9, and criterion-referenced accountability measures such as the CST. Bennicoff-Nan concluded that Star Reading was "a time and labor effective" means of progress monitoring in the

classroom, as well as suitable for program evaluation and monitoring student progress toward state accountability goals.

Linking Star and State Assessments: Comparing Student- and School-Level Data

With an increasingly large emphasis on end-of-the-year summative state tests, many educators seek out informative and efficient means of gauging student performance on state standards—especially those hoping to make instructional decisions before the year-end assessment date.

For many teachers, this is an informal process in which classroom assessments are used to monitor student performance on state standards. While this may be helpful, such assessments may be technically inadequate when compared to more standardized measures of student performance.

Recently the assessment scale associated with Star Reading has been linked to the scales used by virtually every state summative reading or ELA test in the US. Linking Star Reading assessments to state tests allows educators to reliably predict student performance on their state assessment using Star Reading scores. More specifically, it places teachers in a position to identify

- which students are on track to succeed on the year-end summative state test, and
- which students might need additional assistance to reach proficiency.

Educators using Star Reading assessments can access Star Performance Reports that allow access to students' Pathway to Proficiency. These reports indicate whether individual students or groups of students (by class, grade, or demographic characteristics) are likely to be on track to meet a particular state's criteria for reading proficiency. In other words, these reports allow instructors to evaluate student progress toward proficiency and make data-based instructional decisions well in advance of the annual state tests. Additional reports automatically generated by Star Reading help educators screen for later difficulties and progress monitor students' responsiveness to interventions.

An overview of two methodologies used for linking Star Reading to state assessments is provided in the following section.

Methodology Comparison

Recently, Renaissance Learning has developed linkages between Star Reading Scaled Scores and scores on the accountability tests of a number of states. Depending on the kind of data available for such linking, these linkages have

been accomplished using one of two different methods. One method used student-level data, where both Star and state test scores were available for the same students. The other method used school-level data; this method was applied when approximately 100% of students in a school had taken Star Reading, but individual students' state test scores were not available.

Student-Level Data

Using individual data to link scores between distinct assessments is commonly used when student-level data are readily available for both assessments. In this case, the distribution of standardized scores on one test (e.g. percentile ranks) may be compared to the distribution of standardized scores on another test in an effort to establish concordance. Recently, the release of individual state test data for linking purposes allowed for the comparison of Star assessments to state test scores for several states. Star test comparison scores were obtained within an eight-week window around the median state test date (+/–4 weeks).

Typically, states classify students into one of three, four, or five performance levels on the basis of cut scores (e.g. Below Basic, Basic, Proficient, or Advanced). After each testing period, a distribution of students falling into each of these categories will always exist (e.g. 30% in Basic, 25% in Proficient, etc.). Because Star data were available for the same students who completed the state test, the distributions could be linked via equipercentile linking analysis (see Kolen & Brennan, 2004) to scores on the state test. This process creates tables of approximately equivalent scores on each assessment, allowing for the lookup of Star scale scores that correspond to the cut scores for different performance levels on the state test. For example, if 20% of students were "Below Basic" on the state test, the lowest Star cut score would be set at a score that partitioned only the lowest 20% of scores.

School-Level Data

While using student-level data is still common, obstacles associated with individual data often lead to a difficult and time-consuming process of obtaining and analyzing data. In light of the time-sensitive needs of schools, obtaining student-level data is not always an option. As an alternative, school-level data may be used in a similar manner. These data are publicly available, thus making the linking process more efficient.

School-level data were analyzed for some of the states included in the student-level linking analysis. In an effort to increase sample size, the school-level data presented here represent "projected" Scaled Scores. Each Star score

was projected to the mid-point of the state test administrations window using decile-based growth norms. The growth norms are both grade- and subject-specific and are based on the growth patterns of more than one million students using Star assessments over a three-year period. Again, the linking process used for school-level data is very similar to the previously described process—the distribution of state test scores is compared to projected Star scores and using the observed distribution of state-test scores, equivalent cut scores are created for the Star assessments (the key difference being that these comparisons are made at the group level).

Accuracy Comparisons

Accuracy comparisons between student- and school-level data are particularly important given the marked resource differences between the two methods. These comparisons are presented for three states[1] in Table 74, Table 75, and Table 76. With few exceptions, results of linking using school-level data were nearly identical to student-level data on measures of specificity, sensitivity, and overall accuracy. McLaughlin and Bandeira de Mello (2002) employed similar methods in their comparison of NAEP scores and state assessment results, and this method has been used several times since then (McLaughlin & Bandeira de Mello, 2003; Bandeira de Mello, Blankenship, & McLaughlin, 2009; Bandeira et al., 2008).

In a similar comparison study using group-level data, Cronin et al. (2007) observed cut score estimates comparable to those requiring student-level data.

1. Data were available for Arkansas, Florida, Idaho, Kansas, 2Kentucky, Mississippi, North Carolina, South Dakota, and Wisconsin; however, only North Carolina, Mississippi, and Kentucky are included in the current analysis.

Table 74: Number of Students Included in Student-Level and School-Level Linking Analyses by State, Grade, and Subject

State	Grade	Reading	
		Student	School
NC	3	2,707	4,923
	4	2,234	4,694
	5	1,752	2,576
	6	702	2,604
	7	440	2,530
	8	493	1,814
MS	3	3,821	6,786
	4	3,472	7,915
	5	2,915	8,327
	6	2,367	7,861
	7	1,424	6,133
	8	1,108	4,004
KY	3	10,776	2,625
	4	8,885	4,010
	5	7,147	4,177
	6	5,003	2,848
	7	2,572	2,778
	8	1,198	1,319

Table 75: Comparison of School Level and Student Level Classification Diagnostics for Reading/Language Arts

State	Grade	Sensitivity[a]		Specificity[b]		False + Rate[c]		False – Rate[d]		Overall Rate	
		Student	School	Student	School	Student	School	Student	School	Student	School
NC	3	89%	83%	75%	84%	25%	16%	11%	17%	83%	83%
	4	90%	81%	69%	80%	31%	20%	10%	19%	82%	81%
	5	90%	77%	69%	83%	31%	17%	10%	23%	81%	80%
	6	85%	85%	75%	75%	25%	25%	15%	15%	81%	81%
	7	84%	76%	77%	82%	23%	18%	16%	24%	80%	79%
	8	83%	79%	74%	74%	26%	26%	17%	21%	79%	76%
MS	3	66%	59%	86%	91%	14%	9%	34%	41%	77%	76%
	4	71%	68%	87%	88%	13%	12%	29%	32%	79%	79%
	5	70%	68%	84%	85%	16%	15%	30%	32%	78%	78%
	6	67%	66%	84%	84%	16%	16%	33%	34%	77%	77%
	7	63%	66%	88%	86%	12%	14%	37%	34%	79%	79%
	8	69%	72%	86%	85%	14%	15%	31%	28%	79%	80%
KY	3	91%	91%	49%	50%	51%	50%	9%	9%	83%	83%
	4	90%	86%	46%	59%	54%	41%	10%	14%	81%	80%
	5	88%	81%	50%	65%	50%	35%	12%	19%	79%	77%
	6	89%	84%	53%	63%	47%	37%	11%	16%	79%	79%
	7	86%	81%	56%	66%	44%	34%	14%	19%	77%	76%
	8	89%	84%	51%	63%	49%	37%	11%	16%	79%	78%

a. Sensitivity refers to the proportion of correct positive predictions.
b. Specificity refers to the proportion of negatives that are correctly identified (e.g. student will not meet a particular cut score).
c. False + rate refers to the proportion of students incorrectly identified as "at-risk."
d. False – rate refers to the proportion of students incorrectly identified as *not* "at-risk."

Table 76: Comparison of Differences Between Achieved and Forecasted Performance Levels in Reading/ Language Arts (Forecast % – Achieved %)

State	Grade	Student	School	Student	School	Student	School	Student	School
NC		**Level I**		**Level II**		**Level III**		**Level IV**	
	3	–6.1%	–1.1%	2.0%	1.1%	3.6%	–0.8%	0.4%	0.9%
	4	–3.9%	–2.0%	–0.1%	1.3%	4.3%	0.4%	–0.3%	0.2%
	5	–5.1%	–1.9%	–0.7%	2.4%	8.1%	–0.7%	–2.3%	0.2%
	6	–2.1%	0.2%	0.8%	–0.4%	3.2%	–11.5%	–2.0%	11.7%
	7	–6.4%	–0.9%	2.9%	–0.4%	6.3%	–0.7%	–2.8%	2.0%
	8	–4.9%	–3.0%	3.0%	0.4%	5.1%	2.3%	–3.1%	0.3%
MS		**Minimal**		**Basic**		**Proficient**		**Advanced**	
	3	5.2%	14.1%	3.9%	0.5%	–6.1%	–13.4%	–3.0%	–1.2%
	4	5.6%	10.9%	0.2%	–3.1%	–3.0%	–5.9%	–2.8%	–1.8%
	5	4.2%	12.6%	0.4%	–6.7%	–2.7%	–7.2%	–1.9%	1.3%
	6	1.9%	6.2%	2.0%	–1.5%	–3.8%	–7.1%	0.0%	2.4%
	7	5.3%	7.0%	1.1%	–2.8%	–6.3%	–5.3%	–0.2%	1.0%
	8	6.8%	5.5%	–1.7%	–2.8%	–4.6%	–4.3%	–0.5%	1.5%
KY		**Novice**		**Apprentice**		**Proficient**		**Distinguished**	
	3	–3.5%	–1.4%	0.8%	–1.4%	6.4%	3.1%	–3.7%	–0.3%
	4	–0.5%	–0.3%	–2.5%	2.9%	6.8%	–2.1%	–3.9%	–0.5%
	5	–1.6%	1.0%	–2.3%	3.7%	9.1%	–2.9%	–5.3%	–1.8%
	6	–1.5%	1.9%	–3.6%	–1.1%	7.3%	0.0%	–2.3%	–0.8%
	7	–0.9%	0.6%	–2.5%	2.5%	6.6%	–1.7%	–3.3%	–1.4%
	8	–0.1%	1.0%	–5.1%	1.1%	8.1%	–3.0%	–2.9%	0.8%

Classification Accuracy and Screening Data: NCRTI

The National Center on Response to Intervention (NCRTI) is a federally funded project whose mission includes reviewing the technical adequacy of assessments as screening tools for use in schools adopting multi-tiered systems of support (commonly known as RTI, or response to intervention). In the July 2011 review, Star Reading earned strong ratings on NCRTI's technical criteria.

When evaluating the validity of screening tools, NCRTI considered several factors:

- classification accuracy

- validity
- disaggregated validity and classification data for diverse populations

NCRTI ratings include four qualitative labels: convincing evidence, partially convincing evidence, unconvincing evidence, or data unavailable/inadequate. Please refer to Table 77 for descriptions of these labels as provided by NCRTI, as well as the scores assigned to Star Reading in each of the categories. Further descriptive information is provided within the following tables.

Table 77: NCRTI Screening Indicator Descriptions

Indicator	Description	Star Reading Score
Classification Accuracy	Classification accuracy refers to the extent to which a screening tool is able to accurately classify students into "at risk for reading disability" and "not at risk for reading disability" categories (often evidenced by AUC values greater than 0.85).	Convincing Evidence
Validity	Validity refers to the extent to which a tool accurately measures the underlying construct that it is intended to measure (often evidenced by coefficients greater than 0.70).	Convincing Evidence
Disaggregated Validity and Classification Data for Diverse Populations	Data are disaggregated when they are calculated and reported separately for specific subgroups.	Convincing Evidence

Aggregated Classification Accuracy Data

Receiver Operating Characteristic (ROC) Curves as defined by NCRTI:

"Receiver Operating Characteristic (ROC) curves are a useful way to interpret sensitivity and specificity levels and to determine related cut scores. ROC curves are a generalization of the set of potential combinations of sensitivity and specificity possible for predictors." (Pepe, Janes, Longton, Leisenring, & Newcomb, 2004)

"ROC curve analyses not only provide information about cut scores, but also provide a natural common scale for comparing different predictors that are measured in different units, whereas the odds ratio in logistic regression analysis must be interpreted according to a unit increase in the value of the predictor, which can make comparison between predictors difficult." (Pepe, et al., 2004)

"An overall indication of the diagnostic accuracy of a ROC curve is the area under the curve (AUC). AUC values closer to 1 indicate the screening measure

reliably distinguishes among students with satisfactory and unsatisfactory reading performance, whereas values at .50 indicate the predictor is no better than chance." (Zhou, Obuchowski & Obushcowski, 2002)

Brief Description of the Current Sample and Procedure

Initial Star Reading classification analyses were performed using state assessment data from Arkansas, Delaware, Illinois, Michigan, Mississippi, and Kansas. Collectively these states cover most regions of the country (Central, Southwest, Northeast, Midwest, and Southeast). Both the Classification Accuracy and Cross Validation study samples were drawn from an initial pool of 79,045 matched student records covering grades 2–11. The sample used for this analysis was 49 percent female and 28 percent male, with 44 percent not responding. Twenty-eight percent of students were White, 14 percent were Black, and 2 percent were Hispanic. Lastly, 0.4 percent were Asian or Pacific Islander and 0.2 were American Indian or Alaskan Native. Ethnicity data were not provided for 55.4 percent of the sample.

A secondary analysis using data from a single state assessment was then performed. The sample used for this analysis was 42,771 matched Star Reading and South Dakota Test of Education Progress records. The sample covered grades 3–8 and was 28 percent female and 28 percent male. Seventy-one percent of students were White and 26 percent were American Indian or Alaskan Native. Lastly, 1 percent were Black, and 1 percent were Hispanic and, 0.7 percent were Asian or Pacific Islander.

An ROC analysis was used to compare the performance data on Star Reading to performance data on state achievement tests. The Star Reading Scaled Scores used for analysis originated from assessments 3–11 months before the state achievement test was administered. Selection of cut scores was based on the graph of sensitivity and specificity versus the Scaled Score. For each grade, the Scaled Score chosen as the cut point was equal to the score where sensitivity and specificity intersected. The classification analyses, cut points and outcome measures are outlined in Table 78. When collapsed across ethnicity, AUC values were all greater than 0.80. Descriptive notes for other values represented in the table are provided in the table footnote.

Table 78: Classification Accuracy in Predicting Proficiency on State Achievement Tests in Seven States[a]

Statistic[b]	Initial Analysis	Secondary Analysis
	Value	Value
False Positive Rate	0.2121	0.1824
False Negative Rate	0.2385	0.2201

Table 78: Classification Accuracy in Predicting Proficiency on State Achievement Tests in Seven States[a]

Statistic[b]	Initial Analysis		Secondary Analysis	
	Value		Value	
Sensitivity	0.7615		0.7799	
Specificity	0.7579		0.8176	
Positive Predictive Power	0.4423		0.5677	
Negative Predictive Power	0.9264		0.9236	
Overall Classification Rate	0.7586		0.8087	
	Grade	AUC	Grade	AUC
AUC (ROC)	2	0.816		
	3	0.839	3	0.869
	4	0.850	4	0.882
	5	0.841	5	0.881
	6	0.833	6	0.883
	7	0.829	7	0.896
	8	0.843	8	0.879
	9	0.847		
	10	0.858		
	11	0.840		
Base	0.20		0.24	
	Grade	Cut Score	Grade	Cut Score
Cut Point	2	228		
	3	308	3	288
	4	399	4	397
	5	488	5	473
	6	540	6	552
	7	598	7	622
	8	628	8	727
	9	708		
	10	777		
	11	1,055		

a. Arkansas, Delaware, Illinois, Kansas, Michigan, Mississippi, and South Dakota.
b. The false positive rate is equal to the proportion of students incorrectly labeled "at-risk." The false negative rate is equal to the proportion of students incorrectly labeled not "at-risk." Likewise, sensitivity refers to the proportion of correct positive predictions while specificity refers to the proportion of negatives that are correctly identified (e.g., student will not meet a particular cut score).

Aggregated Validity Data

Table 79 provides aggregated validity values as well as concurrent and predictive validity evidence for Star Reading. Median validity coefficients ranged from 0.68–0.84.

Table 79: Overall Concurrent and Predictive Validity Evidence for Star Reading

Type of Validity	Grade	Test	N (Range)	Coefficient	
				Range	Median
Predictive	3–6	CST	1,000+	0.78–0.81	0.80
Predictive	2–6	SAT9	44–389	0.66–0.73	0.68
Concurrent	1–8	Suffolk Reading Scale	2,694	0.78–0.88	0.84
Construct	3, 5, 7, 10	DRP	273–424	0.76–0.86	0.82
Concurrent	1–4	DIBELS Oral Reading Fluency	12,220	0.71–0.87	0.81
Predictive	1–6	State Achievement Tests	74,877–200,929	0.68–0.82	0.79
Predictive	7–12	State Achievement Tests	3,107–64,978	0.81–0.86	0.82
Concurrent	3–8	State Achievement Tests	1,200–2,329	0.71–0.74	0.73
Predictive	3–8	State Achievement Tests	2,974–4,493	0.66–0.70	0.68

Disaggregated Validity and Classification Data

Table 80 shows the disaggregated classification accuracy data for ethnic subgroups and also the disaggregated validity data.

Table 80: Disaggregated Classification and Validity Data

Classification Accuracy in Predicting Proficiency on State Achievement Tests in 6 States (Arkansas, Delaware, Illinois, Kansas, Michigan, and Mississippi): by Race/Ethnicity					
	White, non-Hispanic (n = 17,567)	Black, non-Hispanic (n = 8,962)	Hispanic (n = 1,382)	Asian/Pacific Islander (n = 231)	American Indian/Alaska Native (n = 111)
False Positive Rate	0.3124	0.4427	0.3582	0.1710	0.1216
False Negative Rate	0.3762	0.1215	0.1224	0.2368	0.4054
Sensitivity	0.6238	0.8785	0.8776	0.7632	0.5946
Specificity	0.8676	0.5573	0.6418	0.8290	0.8784
Positive Predictive Power	0.5711	0.5031	0.6103	0.4677	0.7097
Negative Predictive Power	0.8909	0.8999	0.8913	0.9467	0.8125
Overall Classification Rate	0.8139	0.6658	0.7337	0.8182	0.7838

Table 80: Disaggregated Classification and Validity Data

Classification Accuracy in Predicting Proficiency on State Achievement Tests in 6 States (Arkansas, Delaware, Illinois, Kansas, Michigan, and Mississippi): by Race/Ethnicity										
	White, non-Hispanic (n = 17,567)		Black, non-Hispanic (n = 8,962)		Hispanic (n = 1,382)		Asian/Pacific Islander (n = 231)		American Indian/Alaska Native (n = 111)	
AUC (ROC)	Grade	AUC	Grade	AUC	Grade	AUC	Grade	AUC	Grade	AUC
	2	n/a	2	0.500	2	n/a	2	n/a	2	n/a
	3	0.863	3	0.828	3	0.868	3	0.913	3	0.697
	4	0.862	4	0.823	4	0.837	4	0.869	4	0.888
	5	0.853	5	0.832	5	0.839	5	0.855	5	0.919
	6	0.849	6	0.806	6	0.825	6	0.859	6	0.846
	7	0.816	7	0.784	7	0.866	7	0.904	7	0.900
	8	0.850	8	0.827	8	0.812	8	0.961	8	1.000
	9	1.000	9	0.848	9	n/a	9	n/a	9	n/a
	10	0.875	10	0.831	10	0.833	10	n/a	10	n/a
	11	0.750	11	1.000	11	n/a	11	n/a	11	n/a
Base Rate	0.2203		0.3379		0.3900		0.1645		0.333	
Cut Scores	Grade	Cut Score	Grade	Cut Score	Grade	Cut Score	Grade	Cut Score	Grade	Cut Score
	2	228	2	228	2	228	2	228	2	228
	3	308	3	308	3	308	3	308	3	308
	4	399	4	399	4	399	4	399	4	399
	5	488	5	488	5	488	5	488	5	488
	6	540	6	540	6	540	6	540	6	540
	7	598	7	598	7	598	7	598	7	598
	8	628	8	628	8	628	8	628	8	628
	9	708	9	708	9	708	9	708	9	708
	10	777	10	777	10	777	10	777	10	777
	11	1,055	11	1,055	11	1,055	11	1,055	11	1,055

Disaggregated Validity					
				Coefficient	
Type of Validity	Age or Grade	Test or Criterion	n (range)	Range	Median
Predictive (White)	2–6	SAT9	35–287	0.69–0.75	0.72
Predictive (Hispanic)	2–6	SAT9	7–76	0.55–0.74	0.675

References

Allington, R., & McGill-Franzen, A. (2003). Use students' summer-setback months to raise minority achievement. *Education Digest, 69*(3), 19–24.

Bennicoff-Nan, L. (2002). *A correlation of computer adaptive, norm referenced, and criterion referenced achievement tests in elementary reading.* Unpublished doctoral dissertation, The Boyer Graduate School of Education, Santa Ana, CA.

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

Betebenner, D. W. (2010). New directions for student growth models. Retrieved from the National Center for the Improvement of Educational Assessment website: http://www.ksde.org/LinkClick.aspx?fileticket=UssiNoSZks8%3D&tabid=4421&mid=10564

Betebenner, D. W., & Iwaarden, A. V. (2011a). SGP: An R package for the calculation and visualization of student growth percentiles & percenitle growth trajectories [Computer Software manual]. (R package version 0.4-0.0 available at http://cran.r-project.org/web/packages/SGP/)

Betebenner, D. W. (2011b). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. The National Center for the Improvement of Educational Assessment. Retrieved from http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf

Borman, G. D. & Dowling, N. M. (2004). *Testing the Reading Renaissance program theory: A multilevel analysis of student and classroom effects on reading achievement.* University of Wisconsin-Madison.

Bracey, G. (2002). Summer loss: The phenomenon no one wants to deal with. *Phi Delta Kappan, 84*(1), 12–13.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally & Company.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin Company.

Deno, S. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37*(3), 184–192.

Diggle, P., Heagerty, P., Liang, K., & Zeger, S. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

Duncan, T., Duncan, S., Strycker, L., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Holmes, C. T., & Brown, C. L. (2003). *A controlled evaluation of a total school improvement process, School Renaissance.* University of Georgia. Available online: http://www.eric.ed.gov/PDFS/ED474261.pdf.

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

Kirk, R. (1995). *Experimental Design: Procedures for the behavioral sciences* (3rd ed.). New York: Brooks/Cole Publishing Company.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

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

Meyer, B., & Rice, G. E. (1984). The structure of text. In P.D. Pearson (Ed.), *Handbook on reading research* (pp. 319–352). New York: Longman.

Moskowitz, D. S., & Hershberger, S. L. (Eds.). (2002). *Modeling intraindividual variability with repeated measures data: Methods and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Multiple authors. (2002). *Modeling intraindividual variability with repeated measures data: Methods and applications.* In D. S. Moskowitz & S. L. Hershberger (eds.), Mahwah, NJ: Lawrence Erlbaum Associates.

Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). New York: WCB McGraw-Hill.

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Index

A

Access levels, 12
Adaptive Branching, 3, 7, 12
Additional validity evidence, 69
Administering the test, 13
Alternate form reliability, 48
Answering test questions, 6
Area Under the Curve (AUC), 77
ATOS, 18
ATOS graded vocabulary list, 17
AUC. See Area Under the Curve (AUC)
Authentic text passage item specifications, 19

B

Bayesian-modal Item Response Theory (IRT), 41

C

Calibration
    of Star Reading items for use in version 2, 28
    of supplemental items for use in version 4.3, 36
California Achievement Test (CAT), 139
California Standards Test (CST), 165
Common Core State Standards (CCSS), 66
Comparing the Star Reading test with classical tests, 102
Compensating for incorrect grade placements, 113
Comprehensive Test of Basic Skills (CTBS), 139
Computer-adaptive test design, 40
Concurrent validity, correlations with reading tests in England, 70
Conditional standard error of measurement (CSEM), 46
Construct validity, correlations with a measure of reading comprehension, 58, 71
Content development, 14
    ATOS graded vocabulary list, 17
    Educational Development Laboratory's core vocabulary list, 17
Content specification
    Star Reading, 14
    Star Reading Progress Monitoring, 18
Content validity, 58
Conversion tables, 114
Cronbach's alpha, 45, 48
Cross-validation study results, 74
CSEM. See Conditional standard error of measurement (CSEM)

D

Data analysis, 95
Data encryption, 12
Definitions of scores, 99
Description of the program, 1
DIBELS oral reading fluency (DORF), 72, 142, 175
DORF (DIBELS oral reading fluency), 72
Dynamic calibration, 14, 15, 38, 39

E

Educational Development Laboratory, core vocabulary list, 17
EIRF. See Empirical item response functions (EIRF)
Emergent Reader, 134, 136
Empirical item response functions (EIRF), 33, 34
England, 70
Estimated Oral Reading Fluency (Est. ORF), 72, 102
Extended time limits, 11
External validity, 62, 165

G

GE. See Grade Equivalent (GE)
Goal setting, 86, 98
Grade Equivalent (GE), 101, 105, 114, 128
Grade placement, 111
    compensating for incorrect grade placements, 113
    indicating appropriate grade placement, 112
Growth estimates, 98
Growth norms, 97, 168

I

Improvements to the Star Reading test in versions 3 RP and higher, 5
Indicating appropriate grade placement, 112
Individualized tests, 12
Instructional Reading Level (IRL), 103, 104
Internal validity, 59
Investigating Oral Reading Fluency and developing the Est. ORF (Estimated Oral Reading Fluency) scale, 72
Iowa Test of Basic Skills (ITBS), 139
Item calibration, 27
    sample description, 28
    sample description, item difficulty, 32
    sample description, item discrimination, 32
    sample description, item presentation, 30
    sample description, item response function, 32
Item development, 14
    authentic text passage item specifications, 19
    vocabulary-in-context item specifications, 18
Item development specifications
    accuracy of content, 25
    adherence to skills, 22
    balanced items, bias and fairness, 24
    efficiency in use of student time, 24
    item components, 25
    language conventions, 25
    level of difficulty, cognitive load, content differentiation, and presentation, 23
    level of difficulty, readability, 22
    Star Reading, 18
Item difficulty, 32
Item discrimination, 32
Item presentation, 30
Item response function (IRF), 32, 34
Item Response Theory (IRT), 32, 47
    difficulty scale, 28
    difficulty scale, parameter, 36
Item retention, rules, 34
Item specifications
    authentic text passages, 19
    vocabulary-in-context items, 18

K

Kuder-Richardson Formula 20 (KR-20), 48

L

Length of test, 7
Lexile® Measures, 108
    Lexile Framework® for Reading, 110
    of students and books, 109
Linking Star and state assessments, 166
    school-level data, 167
    student-level data, 167
Longitudinal study, correlations with SAT9, 69

M

Maximum likelihood IRT estimation, 27
Measurement precision, 45
Metropolitan Achievement Test (MAT), 139

N

National Center on Response to Intervention (NCRTI), 139, 171, 172
    disaggregated validity and classification data, 78
Normal Curve Equivalent (NCE), 106, 140
Norming, 89
    data analysis, 95
    growth norms, 97

sample characteristics, 90
test administration, 95
test score norms, 89

O

Oral Reading Fluency, 72

P

Password entry, 13
Pathway to Proficiency, 166
Percentile Rank (PR), 105
Permissions, 12
Post-publication studies, 165
Post-publication study data
correlations with a measure of reading comprehension, 71
correlations with reading tests in England, 70
correlations with SAT9, 69
investigating Oral Reading Fluency and developing the Est. ORF (Estimated Oral Reading Fluency) scale, 72
PR. See Percentile Rank (PR)
Practice session, 6
Probable Readers, 134, 136
Program description, 1
Purpose of the program, 1

R

Rasch IRT (Item Response Theory) model, 32
Reader
Emergent, 134, 136
Probable, 134, 136
Transitional, 134, 136
Receiver Operating Characteristic (ROC) curves, 172
Reliability, 45
alternate-form, 48
coefficient, 45, 54
definition, 45
split-half, 45, 48
standard error of measurement (SEM), 51
test-retest, 45, 48
Repeating a test, 9
ROC analysis, 77
Rules for item retention, 34

S

Sample characteristics, norming, 90
SAT9, 69
Scale calibration, 27, 36
Scaled Score (SS), 36, 99
Score definitions, 99
conversion tables, 114
grade placement, 111
special scores, 111
types of test scores, 99
Scores
Estimated Oral Reading Fluency (Est. ORF), 72, 102
Grade Equivalent (GE), 101, 104, 105
Instructional Reading Level (IRL), 103, 104
Lexile® Measures, 108
Normal Curve Equivalent (NCE), 106
Percentile Rank (PR), 105
Scaled Score (SS), 99
special IRL (Instructional Reading Level), 104
Student Growth Percentile (SGP), 107
test scores, 99
Zone of Proximal Development (ZPD), 111
Security. See Test security
SEM. See Standard error of measurement (SEM)
SGP. See Student Growth Percentile (SGP)
Special scores
Zone of Proximal Development (ZPD), 111
Split-application model, 12
Split-half reliability, 45, 48
SS. See Scaled Score (SS)
Standard error of measurement (SEM), 51, 141
Stanford Achievement Test (SAT9), 139, 165
Star Early Learning, 90
Star Reading, compared with classical tests, 102
State assessments, linked to Star, 166
methodology comparison, 166
school-level data, 167
student-level data, 167
State tests, 65, 159

Student Growth Percentile (SGP), 107
Summary of validity data, 87

T

Test administration, 95
Test administration procedures, 13
Test items, time limits, 10
Test monitoring, 13
Test repetition, 9
Test-retest reliability, 45, 48
Test score norms, 89
Test security, 11
    access levels and user permissions, 12
    data encryption, 12
    individualized tests, 12
    split-application model, 12
    test monitoring and password entry, 13
Time limits, 10
    extended time limits, 11
Transitional Reader, 134, 136
Types of test scores, 99
    comparing the Star Reading test with classical tests, 102
    Estimated Oral Reading Fluency (Est. ORF), 72, 102
    Grade Equivalent (GE), 101, 105
    Instructional Reading Level (IRL), 103, 104
    Normal Curve Equivalent (NCE), 106
    Percentile Rank (PR), 105
    Scaled Score (SS), 99
    special IRL (Instructional Reading Level), 104
    Student Growth Percentile (SGP), 107

U

Understanding GE scores, 104
Understanding IRL scores, 104
User permissions, 12

V

Validation evidence
    additional, 69
Validity
    concurrent, 70
    construct, 71
    cross-validation study results, 74
    definition, 58
    external evidence, 62
    internal evidence, 59
    investigating Oral Reading Fluency and developing the Estimated Oral Reading Fluency (Est. ORF) scale, 72
    longitudinal study, 69
    summary of validity data, 87
Vocabulary-in-context item specifications, 18
Vocabulary lists, 17

W

Words correctly read per minute (WCPM), 73

Z

Zone of Proximal Development (ZPD), 111, 128

About Renaissance

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

RENAISSANCE®

© Copyright 2019 Renaissance Learning, Inc. All rights reserved. | (800) 338-4204 | www.renaissance.com

All logos, designs, and brand names for Renaissance's products and services, including, but not limited to, Accelerated Math, Accelerated Reader, Accelerated Reader 360, AccelScan, English in a Flash, MathFacts in a Flash, myON, myON Reader, myON News, Renaissance, Renaissance Flow 360, Renaissance Growth Platform, Renaissance Growth Alliance, Renaissance Learning, Renaissance-U, Renaissance Smart Start, Star, Star 360, Star Custom, Star Early Literacy, Star Early Literacy Spanish, Star Math, Star Math Spanish, Star Reading, Star Reading Spanish, and Star Spanish are trademarks of Renaissance. All other product and company names should be considered the property of their respective companies and organizations.

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

R43843.191112