RENAISSANCE®

Star Assessments™ for Early Literacy Technical Manual

RENAISSANCE
Star
Early Literacy®

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 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 the Star Early Literacy Renaissance software. For more information, contact Renaissance Learning, Inc., at the address above.

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, 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 Learning, Inc. All other product and company names should be considered the property of their respective companies and organizations.

METAMETRICS®, LEXILE®, and LEXILE® FRAMEWORK are trademarks of MetaMetrics, Inc., and are registered in the United States and abroad. Copyright © 2019 MetaMetrics, Inc. All rights reserved.

Macintosh is a trademark of Apple Inc., registered in the U.S. and other countries.

WINSTEPS is a registered trademark of John M. Linacre.

11/2019 SELRP

Contents

Introduction 1

Star Early Literacy: Screening and Progress-Monitoring Assessment 1

Star Early Literacy Purpose 2

Design of Star Early Literacy 3

- Test Interface 4
- Pretest Instructions 4
- Hands-On Exercise 5
- Practice Session 5
- Adaptive Branching/Test Length 5
- Test Repetition 6
- Item Time Limits 7
- Repeating the Instructions 7

Test Security 7

- Split Application Model 8
- Individualized Tests 8
- Data Encryption 8
- Access Levels and Capabilities 8
- Test Monitoring/Password Entry 9
- Final Caveat 9

Content and Item Development 10

Content Specification 10

Test Blueprint Characteristics 11

- Blueprint Sub-Domain Prescriptions 12

The Star Early Literacy Item Bank 12

Item Design Guidelines 18

- Simplicity 18
- Screen Layout 18
- Text 19
- Graphics 19
- Phonics and Phonemic Awareness 19
- Answer Options 21
- Language and Pronunciation 21

Item Development: Star Early Literacy (Prototype Testing) ... 22
Item Development: Star Early Literacy ... 23
Balanced Items: Bias and Fairness ... 23
Content Structure ... 24
Tagging for "Requires Reading" Issue ... 24
Metadata Requirements and Goals ... 25
Text ... 26
Readability Guidelines ... 26

Item and Scale Calibration ... 28

Background ... 28
Calibration of Initial Star Early Literacy Items ... 28
Sample Description ... 29
Item Response Function ... 32
Item Retention ... 34
Score Scale Definition and Development ... 35
Computer-Adaptive Test Design ... 36
Scoring in the Star Early Literacy Tests ... 37
On-line Data Collection for New Item Calibration ... 37
A New Scale for Reporting Star Early Literacy Test Scores ... 38

Reliability and Measurement Precision ... 41

Generic Reliability ... 42
Split-Half Reliability ... 44
Retest (Alternate Forms) Reliability ... 44
Standard Error of Measurement ... 45

Validity ... 48

Relationship of Star Early Literacy Scores to Age and School Grade ... 49
Relationship of Star Early Literacy Scores to Skills Ratings ... 50
The Rating Instrument ... 51
Skills Rating Items Used in the Spring 2012 Star Early Literacy Skills Ratings Survey ... 51
Relationship of Scaled Scores to Skills Ratings ... 52
Relationship of Star Early Literacy Scores to Other Tests ... 53

Validation Study Data 53
Meta-Analyses of the Validation Study Validity Data 56
Evidence of Technical Adequacy for Informing Screening and Progress Monitoring Decisions 57
Screening 58
Progress Monitoring 59
Ad Hoc Studies of Star Early Literacy Validity 63
Running Record 63
Michigan Literacy Progress Profile (MLPP) 64
DIBELS, GRADE, and TPRI 65
Predicting Reading Performance from Star Early Literacy Scores 67
Concurrent Validity of Estimated Oral Reading Score 69
Equivalence and Validity of the Current and Original Star Early Literacy Versions 71
The Validity of Early Numeracy Test Items as Measures of the Early Literacy Construct 73
The Dimensionality of Star Early Literacy 75
Summary of Star Early Literacy Validity Data 79

Norming 81

Background 81
Previous Star Early Literacy Norms 81
The 2017 Star Early Literacy Norms 81
Sample Characteristics 82
Test Administration 86
Data Analysis 86
Growth Norms 88

Score Definitions 90

Types of Test Scores 90
Star Early Literacy Scale Scores 91
Unified Scale Scores 91
Sub-domain and Skill Set Scores 92
Literacy Classification 92
Estimated Oral Reading Fluency (Est. ORF) 93
Percentile Ranks (PR) 93

Normal Curve Equivalent (NCE) Scores 94
Grade Equivalent (GE) 95
Comparing Star Early Literacy with Conventional Tests 96
Student Growth Percentile (SGP) 96
Lexile® Measures and Lexile® Ranges 98

Conversion Tables 99

Appendix: Detailed Validity Data 108

Relationship of Star Early Literacy Scores to Other Tests 108
Calibration Study Results 108
Validation Study and Later Study Data 109

References 115

Index 117

Introduction

Star Early Literacy: Screening and Progress-Monitoring Assessment

Star Early Literacy is a computer-adaptive assessment instrument designed to measure the early literacy skills of beginning readers. Star Early Literacy addresses the need to determine children's mastery of literacy concepts that are directly related to their future success as readers. Star Early Literacy assesses proficiency in three broad domains (Word Knowledge and Skills, Comprehension Strategies and Constructing Meaning, and Numbers and Operations) which include ten key early literacy sub-domains involving 41 different sets of skills or concepts. Star Early Literacy was designed explicitly to be used to assess children in kindergarten through grade 2. However, throughout its research and development, it was administered satisfactorily to children from pre-kindergarten through grade 3. In many cases, it will be suitable for teachers' use in assessing pre-kindergarten students and/or students in grade 3 and beyond.

Early childhood education programs abound in this country. Whether federally funded Head First, Even Start and Head Start programs, public preschools administered by local school districts, or private programs that are typically associated with parochial schools, the importance of assessing early literacy skills cannot be overstated. The continued ability to assess these skills during the early primary grades will enable teachers to intervene early in the formal learning process. Research supports successful early intervention as the single best predictor for future academic success, particularly in the critical areas of reading and language acquisition.

Star Early Literacy is distinguished from other assessments of early literacy in three ways. First, it is computer-administered, requiring a minimum of oversight by the teacher; its use of computer graphics, audio instructions, and computerized, automatic dictation of instructions and test questions means that most children can take the test without teacher assistance. Second, its administration is computer-adaptive, which means the content and difficulty levels of the assessment are tailored to each student's performance. Third, it is brief; each assessment administers just 27 test items and takes an average of less than ten minutes. Despite its brevity, Star Early Literacy has been shown to correlate highly with a wide range of more time-intensive standardized measures of early literacy, reading, and other learning readiness skills.

Unlike many assessments, Star Early Literacy is designed specifically for repeated administration throughout the school year. Star Early Literacy incorporates an automated database that records the results of each assessment and makes them available immediately in reports of the status and growth of individual students and classes as a whole.

Star Early Literacy Purpose

Star Early Literacy is designed to provide teachers with both norm-referenced and criterion-referenced scores that will help in planning instruction and monitoring the progress of each student. Star Early Literacy supports regular assessments on a variety of literacy skills throughout the school year. This will enable teachers to easily track progress and adjust instruction based on students' current needs.

Students are expected to develop a variety of early literacy skills as they progress from pre-kindergarten through third grade. This progression reflects both the home literacy environment and educational interventions. The development of these skills is not, however, continuously upward. Students sometimes learn a skill, forget it, and relearn it, a cycle that is perfectly normal. Many well-established tests are available that test early literacy skills at a point in time, but few are designed to repeatedly assess a child's status at different stages through this important growth period.

Regular assessment can provide teachers with timely information concerning student understanding of literacy concepts and will prove more useful than one-time assessments. Regular assessment will also help teachers determine weekly classroom activities that will introduce students to new skills, provide them with practice so they can improve existing skills, and review skills that students may have forgotten.

Star Early Literacy is designed for regular assessment of literacy skills and concepts in kindergarten through second grade students. In many cases, its use will be appropriate in pre-kindergarten as well as in grade 3 and beyond. Star Early Literacy provides teachers with immediate feedback that will highlight instructional needs and enable teachers to target literacy instruction in order to improve the overall literacy skills of their students by some measurable means.

Star Early Literacy:

- Assesses the early literacy skills of pre-kindergarten through third grade students.

- Identifies specific areas of strength and weakness in the sub-domains and skills assessed by the program.
- Identifies students who may be at risk for later reading failure.
- Provides teachers with the following:
  - information that can be used for goal setting and outcome assessment
  - measurable information regarding individual and class literacy skills
  - timely and accurate information that can be used to plan literacy instruction and intervention
  - a tool that enables them to capture a comprehensive picture of student literacy skills in ten sub-domains
- Helps teachers monitor student progress based on the specific literacy needs of each student.

Design of Star Early Literacy

The 27-item Star Early Literacy test is the latest and only version of Star the Early Literacy assessments. While it improves on previous versions in a number of ways, it also retains many features of the previous versions. One of the fundamental design decisions concerning Star Early Literacy involved the choice of how to administer the test. The primary advantage of using computer software to administer Star Early Literacy tests is the ability to tailor each student's test based on his or her responses to previous items. Paper-and-pencil or fixed tests are obviously far different from this: 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 in Star Early Literacy, effectively customize the test to each student's achievement level.

Adaptive Branching offers significant advantages in terms of test reliability, testing time, and student motivation. Reliability improves over paper-and-pencil or fixed tests because the test difficulty matches 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 fixed 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 design decision concerning Star Early Literacy involved the choice of the content and format of items for the test. Its content spans three domains and ten sub-domains of early literacy skills and abilities, ranging from general readiness to vocabulary, and includes four of the five key areas of reading instruction recommended by the National Reading Panel report: phonemic awareness, phonics, vocabulary, and text comprehension. The format of its test items is engaging to young children, using graphics, animation, and digitized voice to present instructions, practice, and the test items themselves.

For these reasons, Star Early Literacy's test design and item format provide a valid procedure to assess pre-reading skills and sentence and paragraph-level comprehension and to identify a student's literacy classification. Data and information presented in this manual reinforce this.

Test Interface

The test interface for Star Early Literacy was designed to be simple, appealing to young school children, and effective. Every test question begins with dictated instructions by means of digitized audio recordings. Additionally, every question is presented in a graphic display format. The student can replay the instructions at will; instructions will replay automatically after a measured time interval if there is no action by the student. All questions are in multiple-choice format with three response alternatives.

Students select their answers by:

- If using the keyboard, students press one of the three keys (**1**, **2**, or **3**) and then press the **Enter** key (or the **return** key on Macintosh computers).
- If using the mouse, students select their answers by pointing and clicking the mouse.
- If using a tablet, students tap their answer choice; then, they tap **Next**.

Pretest Instructions

Prior to the test session itself, a brief demonstration video introduces Star Early Literacy to the student. It presents instructions on what to expect, how to use the mouse, keyboard, or tablet, and how to answer the multiple-choice test questions.

Hands-On Exercise

To ensure that every student understands how to use the mouse, keyboard, or tablet, a short hands-on exercise precedes the assessment. The tutorial tests one of three abilities:

1. The student's ability to move the mouse pointer to a target, and to click the mouse pointer on the target,
2. The student's ability to press the correct key on the keyboard to choose his or her answer, and to remember to press **Enter** to move on to the next question, *or*
3. The student's ability to tap his or her answer, and to remember to tap **Enter** to move on to the next question.

Students must demonstrate proficiency in using the relevant response device before the test will proceed. A student must correctly respond to three hands-on exercise questions in a row in order to "test out" of the hands-on exercise. To correctly respond to a question, the student must have no more than one incorrect key press or off-target click (not including the Listen button) and must select the target object within five seconds after the audio instructions are through playing. When software detects that the student is having difficulty using the mouse or keyboard, the student will be instructed to ask the teacher for help.

Practice Session

After satisfactory completion of the hands-on exercise, a short practice test precedes the assessment itself. As soon as a student has answered three of five practice questions correctly, the program takes the student into the actual Star Early Literacy test. Even the youngest students should be able to answer the practice questions correctly. If the student has not successfully answered three questions in the first set of five, a second set of five practice questions is presented. Only after the student has passed the practice test does the actual test begin. Otherwise, Star Early Literacy will halt the testing session and tell the student to ask the teacher for help.

Adaptive Branching/Test Length

Star Early Literacy's branching control uses a proprietary approach somewhat more complex than the simple Rasch maximum information Item Response Theory (IRT) model. The Star Early Literacy approach was designed to yield reliable test results by adjusting item difficulty to the responses of the individual being tested while striving to minimize student frustration.

In order to minimize student frustration, the first administration of Star Early Literacy begins with items that have difficulty levels substantially below what a typical student at a given age and grade level can handle. On the average, about 90 percent of students will be able to answer the first item correctly. After the first two items, Star Early Literacy strikes a balance between student motivation and measurement efficiency by tailoring the choice of test items such that students answer an average of 75 percent of items correctly. On the second and subsequent administrations, Star Early Literacy begins testing the student at the level of his or her most recent score, again adjusting the difficulty of the early items to avoid frustration.

Once the testing session is underway, Star Early Literacy administers 27 items of varying difficulty based on the student's responses; this is sufficient information to obtain a reliable Scaled Score and to estimate the student's proficiency in all of the literacy sub-domains assessed. The average length of time to complete a Star Early Literacy test (not including the pretest instructions) is 9 minutes, with a standard deviation of approximately 2 minutes as shown in Table 1 below. Most students will be able to complete a Star Early Literacy test in under 14 minutes, including pretest instructions, and almost all will be able to do so in less than 17 minutes.

Table 1: Average and Percentiles of Total Time in Minutes to Complete Star Early Literacy Assessment During the 2015–2016 School Year

Grade	Sample Size	Mean	Standard Deviation	5th Percentile	50th Percentile	95th Percentile	99th Percentile
Pre-K	116,848	8.2	2.4	5.7	7.6	12.6	16.7
K	2,461,903	8.1	2.2	5.8	7.6	12.1	15.4
1	1,831,579	8.6	2.2	6.0	8.1	12.5	15.4
2	418,238	9.1	2.3	6.2	8.7	13.3	16.3
3	100,706	8.9	2.3	6.1	8.5	13.0	16.1

Test Repetition

Star Early Literacy 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 Early Literacy may be administered as frequently as weekly for progress monitoring purposes.

The Star Early Literacy item bank contains more than 3,400 items, so students can test often without getting the same questions more than once. Star Early Literacy 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 30-day period.

Item Time Limits

The Star Early Literacy test has time limits for individual items that are based on latency data obtained during item calibration. These time limits are imposed not to ensure rapid responses, but to keep the test moving should the student become distracted and to ensure test security should the student walk away. Items that time out are counted as incorrect responses. (If the student selects the correct response, but does not press enter or return by time-out, the item is counted as a correct response.) Students have up to 35 seconds to answer each hands-on exercise question, up to 60 seconds to answer each practice question, and up to 90 seconds to answer each actual test question. When a student has only 15 seconds remaining for a given item (10 seconds during the hands-on exercise), a chime sounds, a clock appears, and the student is reminded to select an answer.

Repeating the Instructions

If a student wants to repeat the instructions for the current item, he or she can do so by pressing the **L** key on the keyboard or clicking/tapping (if using a tablet) the **Listen** button on the screen. This will cause the instructions to be replayed. The instructions will also be replayed automatically if there is no student action within a preset interval following the initial play of the instructions. The length of that interval varies according to item type, with a longer interval in the case of items that require more time for the student to process them.

Test Security

Star Early Literacy includes many features intended to provide adequate security to protect the content of the test and to maintain the confidentiality of the test results.

Split Application Model

In the Star Early Literacy software, when students log in, they do not have access to the same functions that teachers, administrators, and other personnel can access. Students are allowed to test, but they have no other features available in Star Early Literacy; therefore they have no access to confidential information. When teachers and administrators log in, they can manage student and class information, set preferences, register students for testing, and create informative reports about student test performance.

Individualized Tests

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

Data Encryption

A major defense against unauthorized access to test content and student test scores is data encryption. All of the items and export files are encrypted. Without the appropriate decryption codes, it is practically impossible to read the Star Early Literacy 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 and the capabilities the user has been granted in the Renaissance program. Each primary position is part of a user group. There are seven user groups: district administrator, district staff, school administrator, school staff, teacher, parent, and student.

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

By default, each user group is granted a specific set of capabilities. Each capability corresponds to one or more tasks that can be performed in the

program. The capabilities in these sets can be changed; capabilities can also be granted or removed on an individual level.

Since users can be assigned to the district and/or one or more schools (and be assigned different primary positions at the different locations), and since the capabilities granted to a user can be customized, there are many, varied levels of access an individual user can have.

The security of the Star Early Literacy 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 capabilities that they have been granted. Personnel who log in to Renaissance (teachers, administrators, and staff) must enter a user name and password before they can access the data and create reports. Parents must also log in with a user name and password before they can access the Parent Report. Without an appropriate user name and password, personnel and parents cannot use the Star Early Literacy software.

Test Monitoring/Password Entry

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

Final Caveat

While Star Early Literacy can do a lot to provide specific measures of test security, the real line of defense against unauthorized access or misuse of the program is the users' responsibility. Educators need to be careful not to leave the program running unattended and to monitor 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 Early Literacy's security and the quality and the validity of its scores.

Content and Item Development

Content Specification

Star Early Literacy consists of 3,424 operational items that align to a set of early literacy skills derived from exemplary state standards as well as the Common Core State Standards and current research.

The initial 2001 release of Star Early Literacy 1.0 was a 25-item adaptive test of early literacy skills, with the test items aligned to 7 literacy domains. Since the revised version of the program was released as Star Early Literacy Enterprise in 2011, it has been a 27-item test, with the test items aligned to 3 broad blueprint domains which are organized into 10 blueprint sub-domains as follows:

Blueprint Domains:

- Word Knowledge and Skills
- Comprehension Strategies and Constructing Meaning
- Numbers and Operations

Blueprint Sub-Domains:

- Alphabetic Principle
- Concept of Word
- Visual Discrimination
- Phonemic Awareness
- Phonics
- Structural Analysis
- Vocabulary
- Sentence-Level Comprehension
- Paragraph-Level Comprehension
- Early Numeracy

Star Early Literacy has separate content specifications for each grade, pre-K to 3, as well as for each of 5 literacy levels defined by scale score intervals.

Blueprint skill sets and blueprint domains in Star Early Literacy are based on extensive analysis of curriculum materials, state standards, and the CCSS, and has been reviewed by early learning consultants.

Early numeracy content will be specified for all tests. Star Early Literacy explicitly includes specified numbers of early numeracy items at each grade level and literacy classification.

This structure encompasses four of the five critical areas of reading instruction identified by the National Reading Panel and CCSS. The one area not covered fully by Star Early Literacy is fluency, a reading behavior that is best assessed by other means. However, fluency is well-known to be highly correlated with other reading skills, such as comprehension and using context to determine word meaning, both of which are assessed in Star Early Literacy.

Furthermore, the assessment estimates students' oral reading fluency and displays these estimates on certain reports. (See page 93 for information on the Estimated Oral Reading Fluency scores.)

Test Blueprint Characteristics

Every Star Early Literacy assessment consists of items that tap knowledge and skills from as many as ten different literacy sub-domains. The items comprise several sets of skills for each sub-domain, with 41 different sets of skills in all.

Content balancing specifications, known as the test blueprint, ensure that a specific number of items from each blueprint sub-domain are administered in every test. A summary of the test blueprint for Star Early Literacy appears here, followed by a summary table of item counts by grade level, literacy classification, and content sub-domain.

The test blueprint specifies item counts from each blueprint sub-domain.

Each Star Early Literacy test consists of 27 scored items, and a separately-specified number of uncalibrated items.

The test is organized into three sections:

1. Section A consist of 14 early literacy items with relatively short audio play times.
2. Section B consists of 8 early literacy items with longer audio play times.
3. Section C consists of 5 early numeracy items presented at the end of each test.

During a single test, with some exceptions, no more than 3 items are administered from the same blueprint skill set.

Blueprint Sub-Domain Prescriptions

For the first test a student takes during a school year, the number of items administered from each blueprint sub-domain is prescribed by grade (pre-K, K, 1, 2, 3).

Subsequent to that initial test, the prescriptions are governed by bands of scale scores on the previous test. Additionally, restrictions in the software program ensure that questions that require the ability to read are not administered to students below the first grade. The previous section,Content and Item Development" contains a detailed list of the ten literacy sub-domains and the 41 skill sets assessed by Star Early Literacy.

The Star Early Literacy Item Bank

Within each of the three Star Early Literacy blueprint domains, closely related skill sets are organized into blueprint sub-domains. The resulting hierarchical structure is domain, sub-domain, skill set, and specific skill. Each Star item is designed to assess a specific skill within the test blueprint. The test blueprint is structured so that it may stay consistent even as state specific Renaissance Reading Learning Progressions may change, as well as the items testing the related skills. Table 2, Table 3, and Table 4 display the domains, sub-domains, skill sets, and skills.

Table 2: Hierarchical Structure of the Star Early Literacy Item Bank: Word Knowledge and Skills Blueprint Domain

Blueprint Sub-Domain	Blueprint Skill Set	Blueprint Skill
Alphabetic Principle	Alphabetic Knowledge	Recognize lowercase letters
		Recognize uppercase letters
		Match lowercase with uppercase letters
		Match uppercase with lowercase letters
		Distinguish numbers from letters
	Alphabetic Sequence	Identify the letter that comes next
		Identify the letter that comes before
	Letter Sounds	Recognize sounds of lowercase letters
		Recognize sounds of uppercase letters

Table 2: Hierarchical Structure of the Star Early Literacy Item Bank: Word Knowledge and Skills Blueprint Domain

Blueprint Sub-Domain	Blueprint Skill Set	Blueprint Skill
Concept of Word	Print Concepts: Word Length	Identify longest word
		Identify shortest word
	Print Concepts: Word Borders	Identify number of words (2–3)
	Print Concepts: Letters and Words	Differentiate words from letters
		Differentiate letters from words
Visual Discrimination	Letters	Differentiate lowercase letters
		Differentiate uppercase letters
		Differentiate lowercase letters in mixed set
		Differentiate uppercase letters in mixed set
	Identification and Word Matching	Identify words that are different
		Match words that are the same
		Identify words that are different from a prompt
Phonemic Awareness	Rhyming and Word Families	Match sounds within word families (named pictures)
		Match sounds within word families (unnamed pictures)
		Identify rhyming words (named pictures)
		Identify nonrhyming words (named pictures)
	Blending Word Parts	Blend onsets and rimes
		Blend 2-syllable words
		Blend 3-syllable words
	Blending Phonemes	Blend phonemes in (VC) or (CVC) words
		Blend phonemes in single-syllable words
	Initial and Final Phonemes	Determine which word (picture) has an initial phoneme different from a prompt
		Determine which word (picture) has a different initial phoneme
		Match initial phoneme to a prompt (pictures)
		Recognize same final sounds (pictures)
		Determine which word (picture) has a final phoneme different from a prompt
		Determine which word (picture) has a different final phoneme
	Consonant Blends (PA)	Match consonant blend sounds (pictures)

Table 2: Hierarchical Structure of the Star Early Literacy Item Bank: Word Knowledge and Skills Blueprint Domain

Blueprint Sub-Domain	Blueprint Skill Set	Blueprint Skill
Phonemic Awareness (continued)	Medial Phoneme Discrimination	Identify short vowel sounds (pictures)
		Identify and match medial sounds (pictures)
		Distinguish short vowel sounds (pictures)
		Match long vowel sounds (pictures)
		Distinguish long vowel sounds (pictures)
	Phoneme Segmentation	Segment syllables in multisyllable words
		Segment phonemes in single-syllable words
	Phoneme Isolation/Manipulation	Substitute initial consonant (named pictures)
		Substitute initial consonant (unnamed pictures)
		Determine missing phoneme, initial or final
		Substitute initial consonant in a prompt (pictures)
		Substitute final consonant sound in a prompt (unnamed pictures)
		Substitute final consonant (named pictures)
		Substitute final consonant sound (unnamed pictures)
		Substitute vowel sounds (pictures)
Phonics	Short Vowel Sounds	Match short vowel sounds (words)
		Match short vowel sounds to letters
		Decode CVC words
		Recognize short vowel sounds (words)
		Distinguish short vowel sounds (words)
		Decode grade-appropriate words
	Initial Consonant Sounds	Identify initial consonant sound (words)
		Identify letter for initial consonant sound (words and letters)
	Final Consonant Sounds	Match word to a given final consonant sound
		Identify letter for a final consonant sound
	Long Vowel Sounds	Identify long vowel sounds (words)
		Match long vowel sounds to a prompt (words)
		Distinguish long vowel sounds (words)
		Match long vowel sounds to letters

Table 2: Hierarchical Structure of the Star Early Literacy Item Bank: Word Knowledge and Skills Blueprint Domain

Blueprint Sub-Domain	Blueprint Skill Set	Blueprint Skill
Phonics (continued)	Long Vowel Sounds (continued)	Decode and recognize associated spelling patterns with long vowels (C-V-C-e)
		Decode and recognize associated spelling patterns with long vowel open syllables
		Decode and recognize associated spelling patterns with long vowel digraphs (including y as a vowel)
	Variant Vowel Sounds	Identify variant vowel sounds
		Identify variant vowel sounds (words)
		Decode words with variant vowels and recognize associated spelling patterns
	Consonant Blends (PH)	Recognize initial consonant blends (words)
		Distinguish consonant blends (words)
		Recognize word with a consonant blend in a contextual sentence
		Recognize associated spelling patterns of initial consonant blends
		Recognize associated spelling patterns of final consonant blends
	Consonant Digraphs	Identify a consonant digraph in a named word
		Identify a consonant digraph in an unnamed word
		Identify a contextual word containing a consonant digraph
		Identify correct spelling of consonant digraphs in words
	Other Vowel Sounds	Identify diphthong sounds in words
		Decode words with diphthongs and recognize associated spelling patterns
		Identify *r-controlled* vowel sounds in named and unnamed words
		Decode words with r-controlled vowels and recognize associated spelling patterns
	Sound-Symbol Correspondence: Consonants	Substitute initial consonants (words)
		Substitute final consonants (words)
		Substitute final consonant sound (named words)
		Substitute final consonant sound (unnamed words)

Table 2: Hierarchical Structure of the Star Early Literacy Item Bank: Word Knowledge and Skills Blueprint Domain

Blueprint Sub-Domain	Blueprint Skill Set	Blueprint Skill
Phonics (continued)	Word Building	Identify words made by adding an initial consonant (unnamed words)
		Identify words made by adding an additional medial letter (unnamed words)
		Identify words made by adding an additional final letter (unnamed words)
		Identify words built by adding one letter to an audio prompt
	Sound-Symbol Correspondence: Vowels	Substitute vowel sounds (words)
	Word Families/Rhyming	Identify rhyming words (words)
		Identify nonrhyming words (words)
		Identify rhyming words (unnamed answers)
		Identify rhyming words (unnamed prompt and answers)
		Identify nonrhyming words (unnamed prompt and answers)
		Identify onset/rime in named words
		Identify onset/rime in unnamed words
		Identify sounds within word families (named words)
		Identify sounds within word families (unnamed words)
Structural Analysis	Words with Affixes	Use knowledge of common affixes to decode words
	Syllabification	Use knowledge of syllable patterns to decode words
		Decode multisyllable words
	Compound Words	Identify compound words (named words)
		Identify words that are not compounds (named words)
		Identify compound words (unnamed words)
		Identify words that are not compounds (unnamed words)
		Identify correctly formed compounds

Table 2: Hierarchical Structure of the Star Early Literacy Item Bank: Word Knowledge and Skills Blueprint Domain

Blueprint Sub-Domain	Blueprint Skill Set	Blueprint Skill
Vocabulary	Word Facility	Match words to pictures
		Read high-frequency words by sight
		Identify new meanings for common multi-meaning words
		Determine categorical relationships
		Understand position words
		Read grade-level sight words
		Understand multi-meaning words
	Synonyms	Identify synonyms of grade-appropriate words
		Match words with their synonyms (words)
		Identify synonym of a grade-appropriate word in a contextual sentence
		Match words with their synonyms in paragraph context (assisted)
		Match words with their synonyms in paragraph context (unassisted)
	Antonyms	Identify antonyms of words
		Identify antonyms of words in context (assisted)
		Identify antonyms of words in context (unassisted)

Table 3: Hierarchical Structure of the Star Early Literacy Item Bank: Comprehension Strategies and Constructing Meaning Domain

Blueprint Sub-Domain	Blueprint Skill Set	Blueprint Skill
Sentence-Level Comprehension	Comprehension at the Sentence Level	Listen and identify word in context
		Read and identify word in context
Paragraph-Level Comprehension	Comprehension of Paragraphs	Identify the main topic of a text
		Listen to text and answer literal *who, what* questions
		Listen to text and answer *where, when, why* questions
		Read text and answer literal *who, what* questions
		Read text and answer *where, when, why* questions

Table 4: Hierarchical Structure of the Star Early Literacy Item Bank: Numbers and Operations Domain

Blueprint Sub-Domain	Blueprint Skill Set	Blueprint Skill
Early Numeracy	Number Naming and Number Identification	Recognize numbers 0–20
	Number Object Correspondence	Count 1–20
		Recognize ordinal numbers 1st–10th
		Compare sets of up to 5 objects
		Identify the number of 10s in 10, 20, 30, 40, 50, 60, 70, 80, 90
	Sequence Completion	Complete a picture pattern
		Complete a sequence of numbers between 0 and 10 in ascending order
	Composing and Decomposing	Add 1 to a set
		Subtract 1 from a set
		Add numbers with a sum up to 10 (pictures)
		Subtract numbers with a minuend up to 10 (pictures)
	Measurement	Compare sizes, weights, and volumes

Item Design Guidelines

Each of the items was written to the following specifications:

Simplicity

Items directly address the skill in the most straightforward manner possible. Evaluators should have no difficulty deducing the exact nature of the skill set or skill being assessed by the item. Instructions should be explicit, clear, simple, and consistent from one item to the next.

Screen Layout

The testing screen should feel comfortable for the student and teacher. Background colors should be unobtrusive and relatively muted, and text and graphics should stand out clearly against the background. The item background must be the same for all items on the test.

Each item consists of a combination of audio instructions, an on-screen prompt in the form of a cloze stem containing text or graphics, and three answer choices containing letters, words, graphics, or numbers.

Text

For letter and word identification items, the type size should be large, but may become smaller for higher grades. The type size should be tied to items, so that it varies according to the developmental level of a student. Text size should always be consistent from one answer choice to the others so as not to make one answer choice stand out visually.

The student instructions for every Star Early Literacy test item is administered aurally by the computer, so there is no need for printed directions on-screen. For any items that utilize on-screen text, either as part of answer choices or for items that require reading, the type is be a sans-serif font of appropriate size.

Every effort is made to use common words as the target and distractor words in test items.

Graphics

Any art should be easily recognized by students. Color should be functional, as opposed to decorative, and lines should be as smooth as possible. For complex graphics, such as those needed for listening comprehension, line drawings on a light background should be used. The size and placement of the graphics should be consistent throughout.

The art for correct answers and distracters should be consistent in order to avoid introducing an extraneous error source. Answer choices will primarily consist of graphics and text, but sound or animation occasionally will be needed. Art should be acceptable to a broad range of teachers, parents, and students, avoiding controversial or violent graphics of any kind.

Phonics and Phonemic Awareness

For phonemic awareness and phonics items, the 44 phonemes (speech sounds) that make up the English language are used. Phonemes should be depicted by one or more letters enclosed in a beginning and ending forward slash mark, as in Table 5.

Table 5: Phonemes Included in the Phonemic Awareness and Phonics Items

Consonant Phonemes		Vowel Phonemes[a]	
/b/	Big, robber	**Short**	
/d/	Down, called, daddy	/ă/	Am, at, apple, pan, laugh
/f/	Phone, if, differ, cough, half	/ĕ/	Ed, end, bread, many, said, friend
/g/	Go, wiggle, ghost	/ĭ/	In, fill, bit, mist
/h/	Had, whole	/ŏ/	On, cot, doll, top
/j/	Gym, job, edge, gem	/ŭ/	Up, but, touch, come, was, does
/k/	Come, keep, back, chrome		
/l/	Let, fell, ample, label, pupil	**Long**	
/m/	Me, swimmer, dumb, Autumn	/ā/	Able, make, aid, day, they, eight, vein
/n/	No, know, winner, gnaw, pneumatic	/ē/	She, seat, bee, key, piece, many, ceiling
/p/	Pay, apple	/ī/	Find, ride, by, pie, high, height
/r/	Read, write, marry, are, rhyme	/ō/	No, note, soul, boat, low, door
/s/	So, cent, pass, house, castle, screw	/ū/	Unit, use, few, you
/t/	To, fatter, debt, ptomaine		
/v/	Very, give, of	**Blended**	
/w/	We, when, quite, once	/o͞o/	Too, super, do, crew, due, two, soup, shoe
/y/	Yes, yellow	/o͝o/	Look, put, could
/z/	Zoo, has, please, buzz, sneeze	/ou/	Mouse, now, drought
/ku/	Quit (really two phonemes /k/ /w/)	/au/	Haul, talk, draw, water, bought, caught
/ks/	Box, fix (really two phonemes /k/ /s/)	/oy/	Oil, boy
	/c/ is always /k/ or /s/		
Digraphs[b]		**/r/ Influenced**	
/sh/	Show, motion, sure	/ar/	Car, far, star
/th/	Thin (unvoiced)	/er/	Her, fur, sir, work, learn, syrup, dollar
/th/	This (voiced)	/or/	For, ore, oar, pour, poor
/ch/	Much, nature, match	/ear/	Rear, ear, hear
/ng/	Song, think	/air/	Air, hair, pair
/wh/	What, when (/wh/ and /w/ often overlap)		

a. 6 vowel letters are used in 70 different spellings and 20 vowel sounds.
b. Single consonant sounds, two letters.

Answer Options

All items have three answer choices. Only one of the choices is the correct answer. Answer choices are always arranged horizontally. For students using the keyboard as the means of choosing their answer, the answers have small text below each answer choice labeling A, B, and C, moving from left to right.

Distractors are chosen to provide the most common errors in recognition, matching, and comprehension tasks.

Words and artwork used in answer choices should be reused in no more than 10% of the items within a skill set, a sub-domain, or within the item bank as a whole. For example, a picture of a cat should only appear as an answer choice in no more than 10 out of 100 items in any given skill set.

Language and Pronunciation

All item directions will be delivered auditorily so as to support students that are not yet reading or emergent readers. Instructions will clearly state the item stem with the instruction for the student to "pick the…" Vocal talent should read questions with a neutral, un-accented American English voice. Pacing should be slow enough to be clearly understood while still be engaging for students.

Language should be used consistently throughout the assessment. Standard protocols should be established for item administration that reflect consistent instructions. For example, if an item stem is repeated twice, the same repetition should be used for all items of the same type. One exception to this rule is those situations where the same item type is used across grades, and one of the factors that changes is the level of instruction provided to the student.

In Phonemic Awareness items, words should be segmented into phonemes, that is, divided into their individual sounds. As much as possible, the individual sounds should be preserved, and not distorted in any way. In the item instructions, individual phonemes will be enclosed by two slash marks, as shown in Table 5.

In the recording of item instructions and answer sound, the audio segments should minimize the tendency to add a vowel sound after a consonant sound, especially for unvoiced consonants, such as /p/, /k/, and /t/. For example, /p/ should not be pronounced "puh." Instead, it should be spoken in a loud whisper and in a clipped manner.

For voiced consonants that cannot be pronounced without a vowel sound, such as /b/ and /g/, the audio segments should keep the vowel sound as short as possible. For example, /g/, not /guh/.

Constituent consonants, such as /m/, /f/, and /n/, should not be followed by a vowel sound. They can, however, be extended slightly, as in /mmmmm/, but not /muh/.

Short and long vowel sounds should be pronounced by simply lengthening the sound of the vowel. The long a sound, for example, should be pronounced /āāāā/.

Item Development: Star Early Literacy (Prototype Testing)

Because Star Early Literacy is intended for computer-administered assessment of early literacy skills of pre-K to grade 3 children who may have limited reading ability, a prototype of the original test delivery software system was developed prior to full-scale item development to evaluate whether this goal was feasible. As part of the product development of Star Early Literacy, prototype test items were written and prototype test administration software was developed, following the guidelines in the previous section. Tryout research of the prototype was carried out in April 2000, with over 1,500 children in pre-kindergarten, kindergarten, and grades 1 and 2 participating. The specific objectives were the following:

- Measure and compare the ability of pre-kindergarten, kindergarten, first, and second grade students to respond to a set of early literacy items, representative of the overall skill set, administered non-adaptively on the computer.
- Gather observations and comments from teachers on the user interface, the overall test, and on individual items as students worked through the test.
- Collect data on how well students interact with the user interface, and determine criteria for testing out of hands-on exercise, repeating instructions, putting up “Get Help” alerts, and other design issues.
- Gather item statistics (percent correct, response latency, amount of mouse travel for students using the mouse, etc., by item and by age/grade) on sets of early literacy items containing text, sound, and graphics.

Extensive analyses were conducted on the data collected in the prototype study to evaluate the software, its user interface, and the psychometric characteristics and teacher opinions of the test items. The results indicated that the prototype tryout study was a success in terms of demonstrating the viability of the software prototype and of the tryout items in classrooms ranging from pre-kindergarten through grade 2.

The user interface proved to be usable at all levels. The tasks were well within the ability of children to complete in a minimum of time. The tryout test items

demonstrated promising psychometric properties. And teachers generally reacted well to the content and format of the prototype. Weak points that were found in the analysis of the tryout study data were corrected in the revised versions of the software used in subsequent studies. (Most weak points were related to correctable audio problems.)

With the blueprint as a guide, items were then written and designed to target the minimum grade level and up for each domain and skill set. For example, an item written at the kindergarten level might include named pictures as answer choices. The same item might then be targeted at the first grade level by using named words as answer choices, and at the second grade level by using unnamed words as answer choices. A total of 2,991 test items were written, spanning the seven domains and 41 skill sets.

Once the test design was determined, individual test items were assembled for tryout and calibration. The item calibration included a total of 2,929 items. It was necessary to write and test about 1,000 questions at each of three grade levels (kindergarten through grade 2) to ensure that at least 600 items per level would be acceptable for the final item collection. Having a pool of almost 3,000 items allowed significant flexibility in selecting only the best items from each domain and skill set for the final product.

Item Development: Star Early Literacy

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.

Content Structure

Every Star Early Literacy assessment consists of 27 items selected adaptively from a bank of 3,424 multiple-choice items (as of January 2018) that tap knowledge and skills from 9 early literacy blueprint sub-domains and 1 blueprint domain for Early Numeracy. Each item contains one correct answer and two distractors.

The Star Early Literacy test administers the same number of items (27) and a separately specified number of uncalibrated items to all students. The items from each sub-domain comprise several skill sets for each sub-domain, with 36 early literacy skill sets and 5 early numeracy skill sets in all. Each time a student at any level takes a Star Early Literacy test, a content-balancing blueprint ensures that a prescribed number of items from each blueprint sub-domain are administered. The number of items per blueprint sub-domain varies by grade and by score ranges on previously taken tests.

There are 4 hierarchical levels: Blueprint Domain, Blueprint Sub-domain, Blueprint Skill Set, and Blueprint Skill

Blueprint Domains: There are 2 early literacy domains (Word Knowledge and Skills; Comprehension Strategies and Constructing Meaning) and 1 early numeracy domain (Numbers and Operations).

Blueprint Sub-domains: There are 9 early literacy sub-domains and 1 early numeracy sub-domain.

Blueprint Skill Sets: There are 36 early literacy skill sets and 5 early numeracy skill sets.

Blueprint Skills: There are 133 early literacy skills and 12 early numeracy skills.

The test itself is organized into three sections:

1. Section A consists of 14 early literacy items with relatively short audio play times.
2. Section B consists of 8 early literacy items with longer audio play times.
3. Section C consists of 5 early numeracy items presented at the end of each test.

Tagging for “Requires Reading” Issue

Some items that require reading may be designated for administration to kindergarten students; versions of Star Early Literacy prior to 2012 have not administered any reading items below grade 1. In order to implement this use

of designated grade K reading items, it is necessary to flag such items for the application software. A single additional grade-like field has been added to indicate the minimum grade an item is to be administered to.

Item selection is filtered by our grade-use rules (currently the maximum item grade is 1 higher than the student's grade) and then further filtered by the minimum allowed student grade.

For example, a grade K item that is not appropriate for Pre-K is marked as:

- Item Grade: K
- Minimum Student Grade: K

A grade K item that can be used for Pre-K students is marked as:

- Item Grade: K
- Minimum Student Grade: Pre-K

Star Early Literacy differs from the other two Star tests with respect to the repetition interval. Star Reading items are used only once in a 90-day interval. Star Math items are used only once in a 75-day interval. For Star Early Literacy, that interval is just 30 days.

Metadata Requirements and Goals

Due to the restrictions for modifying text, the content may not meet the following goals; however, new item development works to bring the content into alignment with these goals:

- **Gender:** After removing gender-neutral items, an equal number of male and female items should be represented. In addition to names (Sara) and nouns (sisters), gender is also represented by pronoun (she). Gender is not indicated by subject matter or appeal. For instance, an item on cooking is not female unless there is a female character in it.
- **Ethnicity:** The goal is to create a balance among the following designations for US products: 60% White, 10% Black or African American, 10% Hispanic, 10% Middle Eastern, and 10% Asian or Indian. Ethnicity can be based on name or subject matter.
- **Subject:** A variety of subject areas should be present across the items, such as Arts/Humanities, Science, History, Physical Education, Math, and Technology.

Metadata is tagged with codes for Genres, Ethnicity, Occupations, Subjects, Topics, and Regions.

Text

Content for Star Early Literacy approximately covers a range of items broad enough to test students from pre-kindergarten through grade 3 as well as remedial students in grade 4. The final collection of test items is large enough so that students can be assessed ten times a year or more without being given the items twice within any 30-day period. There are also enough test items for assessing skills in ten sub-domains. The following sub-domains are considered essential in reading development:

1. **Alphabetic Principle (AP)—**Knowledge of letter names, alphabetic letter sequence, and the sounds associated with letters.
2. **Concept of Word (CW)—**Understanding of print concepts regarding written word length and word borders and the difference between words and letters.
3. **Visual Discrimination (VS)—**Differentiating both upper- and lowercase letters, identifying words that are different and matching words that are the same.
4. **Phonemic Awareness (PA)—**Understanding of rhyming words, ability to blend and segment word parts and phonemes, isolating and manipulating initial, final, and medial phonemes, and identifying the sounds in consonant blend.
5. **Phonics (PH)—**Understanding of short, long, variant vowels, and other vowel sounds, initial and final consonants, consonant blends and digraphs, consonant and vowel substitution, and identification of rhyming words and sounds in word families.
6. **Structural Analysis (SA)—**Understanding affixes and syllable patterns in decoding, and identification of compound words.
7. **Vocabulary (VO)—**Knowledge of high-frequency words, regular and irregular sight words, multi-meaning words, and words used to describe categorical relationships, position words, synonyms, and antonyms.
8. **Sentence-Level Comprehension (SC)—**Identification of words in context.
9. **Paragraph-Level Comprehension (PC)—**Identification of the main topic of text and ability to answer literal and inferential questions after listening to or reading text.

Readability Guidelines

ATOS GLE and word counts are tracked in metadata and used as guides for developing content. The readability levels for each script within each item

should not exceed the grade level of the Star Early Literacy item. Words used in scripts should be appropriate for the intended grade.

The content in Star Early Literacy is leveled to address pre-readers and beginning readers (generally children of ages 3 through 9).

Items in each of the sub-domains were designed to range from easy to difficult. This was achieved through the use of different combinations of audio and graphic elements, such as named pictures, unnamed pictures, named letters and sounds, unnamed letters and sounds, named words, and unnamed words, sentences, and paragraphs. The level of difficulty for each question was controlled through the use of graphical, textual, and audio support.

Generally text at this level follows the following guidelines (which applied to the older version of Star Early Literacy and still applies to the current version of Star Early Literacy):

Table 6: Maximum Sentence Length/Grade

Item Year Group	Maximum Sentence Length[a]
Pre-Kindergarten–Grade 1	10 words
Grades 2–3	12 words

a. Including missing word blank.

Every effort should be made to present age-appropriate vocabulary as well.

The design of items for new skills followed this process:

- Research into current practice and theory was undertaken.
- An expert with an assessment background in early literacy was engaged to provide guidance on the development.
- The Media team provided input on the wording of scripts to ensure consistency with other items in the product.
- After new items were created, they went through the calibration process, after which they were analyzed for their effectiveness by psychometricians and the Content team.

Item and Scale Calibration

Background

Star Early Literacy was initially published in 2001. For the initial version, some 3,000 items measuring pre-literacy skills were developed in calibrated. The initial items were calibrated from response data collected using conventional means: multiple fixed field test forms. Since then, a much more powerful and efficient means of collecting response data on new items has been introduced: dynamic calibration, in which new items are administered on line, embedded as unscored items within operational Star Early Literacy tests. This chapter summarizes the process of item and scale calibration using both methods. It also introduces a new score scale, the Unified Scale, which was introduced in the 2017–18 school year.

Calibration of Initial Star Early Literacy Items

In the course of developing the item bank for the initial version of Star Early Literacy, item writers wrote almost 3,000 test items to measure early literacy skills. This section describes the process by which those items were calibrated on a common scale of difficulty. A later section will describe the process that has been used subsequently to calibrate new items; we call that process "dynamic calibration." Subject matter experts and editors reviewed the content of every item and recommended retaining some and rejecting others. After this item content review, 2,929 items, measuring seven broad literacy areas and 41 literacy-related skills, remained as candidates for inclusion in the item bank.

In order to use the test items for computer-adaptive testing, every item had to be placed on a continuous scale of difficulty—the same scale used to select items adaptively and to score the adaptive tests. The procedures of IRT were chosen as the basis for scaling Star Early Literacy item difficulty, a process called "calibration."

IRT calibration is based on statistical analysis of response data—it requires hundreds of responses to every test item. To obtain these data, Renaissance Learning conducted a major item Calibration Study in late 2000.

Sample Description

A sample of 32,493 students in pre-kindergarten through grade 3 in over 300 schools in the United States participated in the Calibration Study. Over 46,000 computer-administered calibration tests were administered. The calibration sample did not need to be nationally representative, but it did require a wide range of student abilities at each grade or age level. Candidate schools for the recruitment mailing were selected from the MDR (Market Data Retrieval) database based on the availability of grades in the grade span of pre-kindergarten through grade 3. These candidate schools were set up in a recruitment matrix or database, and segmented into cells by geographic region, per-grade district enrollment, and socioeconomic status information.

Table 7 compares certain sample characteristics of the students participating in the Calibration Study against national percentages of the same characteristics.

Table 7: Sample Characteristics, Star Early Literacy Calibration Study, Fall 2000 (N = 32,493 Students)

		Students	
		National %	**Sample %**
Geographic Region	Northeast	20.4	7.8
	Midwest	23.5	21.8
	Southeast	24.3	41.5
	West	31.8	28.9
District Socioeconomic Status	Low	28.4	30.9
	Average	29.6	43.4
	High	31.8	16.3
	Non-Public	10.2	9.4
School Type and District Enrollment	Public		
	< 200	15.8	24.3
	200–499	19.1	23.0
	500–1,999	30.2	29.1
	> 1,999	24.7	14.2
	Non-Public	10.2	9.4

In addition to the sample characteristics summarized in Table 7, additional information about participating schools and students was collected. This information is summarized in Table 8, Table 9, and Table 10. These tables also include national figures based on 2001 data provided by MDR.

Table 8: School Locations, Star Early Literacy Calibration Study, Fall 2000 (N = 308 Schools, 32,493 Students)

	Schools		Students	
	National %	Sample %	National %	Sample %
Urban	27.8	24.7	30.9	23.4
Suburban	38.3	31.2	43.5	31.7
Rural	33.2	43.8	24.8	44.6
Unclassified	0.7	0.3	0.7	0.3

Table 9: Nonpublic School Affiliations, Star Early Literacy Calibration Study, Fall 2000 (N = 36 Schools, 3,056 Students)

	Schools		Students	
	National %	Sample %	National %	Sample %
Catholic	39.7	68.6	51.8	72.3
Other	60.3	31.4	48.2	27.7

Table 10: Ethnic Group Participation, Star Early Literacy Calibration Study, Fall 2000 (N = 32,493 Students)

		Students	
		National %	Sample %
Ethnic Group	Asian	3.4	0.7
	Black	14.5	9.3
	Hispanic	12.7	6.7
	Native American	0.9	0.4
	White	54.7	38.8
	Unclassified	13.8	44.0

Recruitment letters and applications were sent to all the candidate schools in the matrix. The response rate was monitored and additional follow-up was conducted as needed to ensure that the calibration sample met minimum student number requirements per grade.

The objectives of the Calibration Study were to:

- Collect sufficient response data to allow IRT item parameters to be estimated for all 2,929 Star Early Literacy items.
- Conduct preliminary research into the psychometric reliability of Star Early Literacy tests, using a test-retest design.
- Assess the degree of relationship between Star Early Literacy scores and a standardized reading achievement test.

In support of the first objective, provisions were made during forms design to facilitate expressing all IRT item parameters on a common scale. To that end, some of the test items were used as "anchor items"—items common to two or more forms that are used to facilitate linking all items to the common scale. Two kinds of anchoring were used: 1) horizontal (form-to-form) anchoring, and 2) vertical (level-to-level) anchoring.

Horizontal anchoring: The purpose of horizontal anchoring is to place all items at a given level on the same scale, regardless of differences among the forms at that level. To accomplish that, several items appeared in all forms at a given level. These horizontal anchor items were chosen to be representative of the seven content domains and to be appropriate for the grade level.

Vertical anchoring: The purpose of vertical anchoring is to place items at adjacent levels on the same scale. To accomplish that, a number of items were administered at each of two adjacent levels: A and B, or B and C. As much as possible, the vertical anchor items were chosen to be appropriate at both the lower and higher levels at which they were used.

Table 11 depicts the distribution of the three types of items within Star Early Literacy calibration test forms. The distribution differs from one level to another. The three item types are horizontal anchor items, vertical anchor items, and unique (non-anchor) items.

Table 11: Number of Anchor Items and Unique Items in Each 40-Item Test Form, by Level

Item Type	Level A Pre-K & K	Level B Grade 1	Level C Grades 2 & 3
Horizontal anchor items	5	7	5
Vertical anchor items	5	11	6
Unique items	30	22	29
Total	40	40	40

For reliable IRT scale linking, it is important for anchor items to be representative of the content of the tests they are used to anchor. To that end, the distribution of anchor items was approximately proportional to the distribution of items among the domains and skills summarized in “Content and Item Development” on page 10.

To accomplish the second objective of the Calibration Study, many of the participating students were asked to take two Star Early Literacy tests so that the correlation of their scores on two occasions could be used to evaluate the retest reliability of Star Early Literacy tests over a short time interval. Topics related to reliability are described in ”Reliability and Measurement Precision” on page 41.

To accomplish the third objective, a subsample of the grade 1, 2 and 3 students also took a computer-adaptive Star Reading 2.x assessment to provide a basis for evaluating the degree of correlation between Star Early Literacy and reading ability. Statistical results are presented in”Validity” on page 48.

Because all Star Early Literacy test items include computerized graphics and audio, the calibration test forms were all computer-administered.

Item Response Function

With the response data from the Calibration Study in hand, the first order of business was to calibrate the items and score the students’ tests. This was done using the “Rasch model,” an IRT model that expresses the probability of a correct answer as a function of the difference between the locations of the item and the student on a common scale.

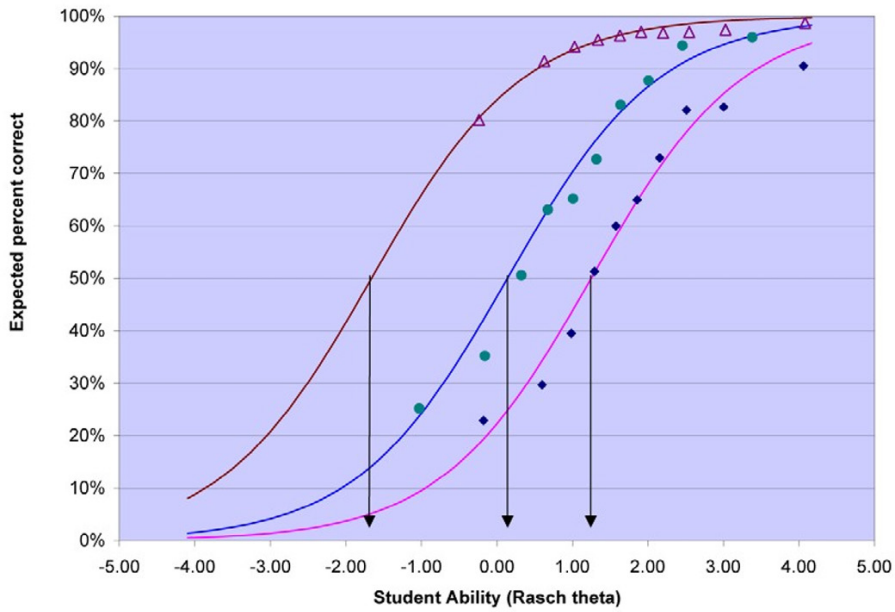
Although IRT encompasses a family of mathematical models, the one-parameter (or Rasch) IRT model was selected for the Star Early Literacy data both for its simplicity and its ability to accurately model the performance of the Star Early Literacy items.

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.

Figure 1: Example of Item Statistics Database Presentation of Information



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.

Rasch model analysis was used to determine the value of a "difficulty parameter" for every item, and to assign a score to every student. In the analysis, a number of statistical measures of item quality and model fit were calculated for each item.

Item parameter estimation and IRT scoring were accomplished using WINSTEPS, a commercially available Rasch model analysis software package. WINSTEPS is capable of Rasch analysis of multiple test forms simultaneously. Using this capability, three item parameter estimation analyses were conducted. All Level B test forms were analyzed first, and the resulting scale was used as the reference scale for the other forms. Following that, separate analyses were conducted of the Level A and Level C forms. In each of the last two analyses, the parameters of anchor items common to Level B were held fixed at the values obtained from the Level B analysis. This had the effect of placing all Level A and Level C item parameters on the Level B scale.[1]

The principal end products of the item calibration process were the IRT item parameter estimates themselves, along with traditional indices of item difficulty (sample proportion correct) and item discriminating power (correlation coefficients between item score and the Rasch ability score).

Item Retention

Once the calibration analysis was complete, a psychometric review took place. The review evaluated both the IRT-based results and the traditional item analysis results, such as proportion correct and item-total correlations.

Reviewers evaluated each item's difficulty, discriminating power, model fit indices, statistical properties and content to identify any items that appeared unsuitable for inclusion in the adaptive testing item bank. The review work was aided by the use of interactive psychometric review software developed specifically for Star Early Literacy. This software displays, one item at a time, the Star Early Literacy question (including audio and graphics) and its correct answer, along with a variety of item statistics. The statistics include

1. All 246 test forms contained a number of anchor items. At each of the three levels, a small set of items specific to that level was common to all of the forms; these "horizontal anchors" served to link all forms at a given level to a common scale. Additionally, every form contained some items in common with forms from adjacent levels; these "vertical anchors" served to link the scales of Levels A and C to the reference scale based on Level B.

Rasch model fit indices, traditional proportion correct and biserial statistics to assess difficulty and discriminating power, an analysis of each response alternative, and the Rasch item difficulty parameter.

There are check boxes for the reviewer to record disqualifying properties and to recommend acceptance, and an area for the reviewer to use to record notes about the item. All reviewers' recommendations and notes were compiled into a permanent database of the psychometric history of all test items developed for use in Star Early Literacy.

Following completion of the psychometric reviews by individual reviewers, a second review of the database was conducted. In that review, differences in reviewer recommendations were reconciled, and final decisions were made about retention or rejection of each item. Of the 2,929 items in the calibration item bank, 2,485 were accepted by the psychometric review team for use in the adaptive version of Star Early Literacy. Of these, 18 were reserved for use as practice items; another 30 items designed specifically as mouse training items were reserved for that purpose. Prior to release of the publication version of Star Early Literacy, a number of other items were deleted in response to independent reviewers' suggestions. The final version of the Star Early Literacy item bank therefore contained 2,369 items.

Score Scale Definition and Development

After item calibration using the Rasch IRT model, a score scale was developed for use in reporting Star Early Literacy results. Although the Rasch ability scale could be used for this purpose, a more "user-friendly" scale was preferred.[2] A system of integer numbers ranging from 300 to 900 was chosen as the score reporting scale for Star Early Literacy.

This scale was chosen in such a way that the range of Scaled Scores is 100 times the age range for which Star Early Literacy was designed—from about 3 to 9. Scaled Score values are very roughly indicative of the typical age of students with similar performance. For example, a Scaled Score of 500 might be expected of 5-year-old students, but would be unexpected among 8-year-olds. Similarly, a Scaled Score of 800 might be expected of 8-year-olds, but would be unusual among 5-year-olds. Scores of 300 and 900, although possible, occur rarely.

2. Scores on the Rasch ability scale are expressed on the "real number" line, use decimal fractions, and can be either negative or positive. While useful for scientific and technical analysis, the Rasch ability scale does not lend itself to comfortable interpretation by teachers and lay persons.

Computer-Adaptive Test Design

In computer-adaptive tests like the Star Early Literacy 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 early literacy skills may branch to easier items in order to better estimate his or her early literacy achievement level.

High-performing students may branch to more challenging early literacy items in order to better determine the breadth of their early literacy skills and their early literacy achievement level.

During a Star Early Literacy test, a student may be "routed" to items at the lowest early literacy level or to items at higher early literacy 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 Early Literacy item calibration studies.

Students who have not taken a Star Early Literacy 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 Early Literacy test within the last six months, the difficulty of the first item depends on that student's previous Star Early Literacy test score information. After the administration of the initial item, and after the student has entered an answer, Star Early Literacy 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 early literacy ability.

Randomization of items with difficulty values near the student's adjusted early literacy ability allows the program to avoid overexposure of test items. Items that have been administered to the same student within the past 30-day 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 Early Literacy computer-adaptive test.

Scoring in the Star Early Literacy Tests

Following the administration of each Star Early Literacy 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 Early Literacy 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 Early Literacy 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 Early Literacy 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 Early Literacy (grades Pre-K–3). 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.

On-line Data Collection for New Item Calibration

Since the initial publication of Star Early Literacy, item development happens on an ongoing basis to continually grow the Star Early Literacy item bank.

Beginning in 2010, "dynamic calibration" has replaced the previous method of collecting and analyzing response data on new Star Early Literacy items. Dynamic calibration allows response data on new test items to be collected during the Star testing sessions for the purpose of field testing and calibrating those items. When dynamic calibration is active, it works by embedding one or more new items at random points during a Star test. These items do not count towards the student's Star test score, but item responses are stored for later psychometric analysis. Students may take as many as five additional items per test; in some cases, no additional items will be administered. On average, this will only increase testing time by one to two minutes. The new, non-calibrated items will not count towards students' final scores, but will be analyzed in conjunction with the responses of hundreds of other students.

Student identification does not enter into the analyses; they are statistical analyses only. The response data collected on new items allows for continual evaluation of new item content and will contribute to continuous improvement in Star tests' assessment of student performance.

A New Scale for Reporting Star Early Literacy Test Scores

Renaissance released Star Early Literacy in 2001, five years following the publication of Star Reading Version 1. As already stated elsewhere, Star Early Literacy is 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 fairly 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). The Rasch transformation formula developed is as follows:

Star Reading Rasch Score = 1.5572 (SEL Rasch Score) – 5.2275

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. A Star Reading 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 reported unified scale is just under 1400; but there is no theoretical limit: If SR content were to be 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.

Table 12 contains a table of selected Star Early Literacy Rasch ability scores and their equivalents on the Star Early Literacy and Unified Score scales.

Table 12: Some Star Early Literacy Rasch Scores and their Equivalents on the Star Early Literacy and Unified Score Scales

Star Early Literacy Rasch Score	Star Early Literacy Scale Score	Unified Scale Score
–8.0626	300	200
–7.2952	301	250
–6.5277	302	300
–5.7603	304	350
–4.9929	308	400
–4.2254	314	450
–3.458	326	500
–2.6906	347	550
–1.9232	383	600
–1.1557	439	650
–0.3883	516	700
0.3791	608	750
1.1466	699	800
1.914	773	850
2.6814	825	900
3.4488	857	950
4.2163	876	1000
4.9837	887	1050
5.7511	893	1100

Reliability and Measurement Precision

Reliability is a measure of the degree to which test scores are consistent—either internally, or across repeated administrations of the same or similar tests to the same group or population. To the extent that a test is reliable, its scores are free from errors of measurement. In educational assessment, however, some degree of measurement error is inevitable. One reason for this is that a student's performance may vary from one occasion to another. Another reason is that variation in the content of the test from one occasion to another may cause scores to vary.

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

Star Early Literacy provides two ways to evaluate the reliability of its scores: reliability coefficients, which indicate the overall precision of a set of test scores; and standard errors of measurement, which provide an index of the degree of error in an individual test score. A reliability coefficient is a summary statistic that reflects the overall amount of measurement precision in a specific examinee group or population as a whole. Standard error of measurement (SEM) estimates the degree of error in the scores themselves. In Star Early Literacy, two different kinds of SEM are calculated: One global, and the other conditional. The conditional standard error of measurement (CSEM) is an estimate of the imprecision of each individual test score. The global standard error is, in effect, an average of the conditional SEM values over a large number of tests. Both the reliability coefficient and the global standard error are single values that apply to the overall test; in contrast, the magnitude of the CSEM may vary substantially from one person's test score to another.

This section presents reliability coefficients of three different kinds: generic reliability, split-half, and test-retest, followed by statistics on the standard error of measurement of Star Early Literacy test scores. Both generic reliability and split-half reliability are estimates of the internal consistency reliability of a test. The reliability coefficients and conditional standard errors of measurement are presented for scores expressed on both the Star Early Literacy Scale and the newly developed Star Unified Scale.

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 the test scores. In Star Early Literacy 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 Early Literacy 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 Early Literacy data from the 2015–2016 school year resulted in the generic reliability estimates shown in Table 13 and Table 14 based on the Unified scale and the Star Early Literacy scale, respectively. Table 13 shows that the generic reliability estimates on the Unified scales range from 0.81 in Pre-K to 0.87 in grade 3. The overall generic reliability is 0.91. The generic reliability estimates based on the Star Early Literacy scale are shown in Table 14 and are very similar to the already described generic reliabilities on the Unified scale. These estimates show a high degree of reliability of the Star Early Literacy scores. 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 Early Literacy.

Table 13: Internal Consistency, Retest Reliability, and Split Half Reliability of Star Early Literacy on the Unified Scale (Assessments Taken in the 2015–2016 School Year)

	Reliability Estimates						
	Generic		Split-Half		Retest		
Grade	N	ρxx	N	ρxx	N	ρxx	Average Days between Testing
Pre-K	90,000	0.81	19,720	0.83	19,500	0.61	101
K	90,000	0.83	19,413	0.84	19,500	0.62	81
1	90,000	0.82	19,129	0.83	19,500	0.58	77
2	90,000	0.84	18,444	0.86	19,500	0.78	77
3	90,000	0.87	17,773	0.90	19,500	0.85	77
Overall	450,000	0.91	94,479	0.92	97,500	0.85	82

Table 14: Internal Consistency, Retest Reliability, and Split Half Reliability of Star Early Literacy on the Enterprise Scale (Assessments Taken in the 2015–2016 School Year)

	Reliability Estimates						
	Generic		Split-Half		Retest		
Grade	N	ρxx	N	ρxx	N	ρxx	Average Days between Testing
Pre-K	90,000	0.83	19,720	0.83	19,500	0.61	101
K	90,000	0.83	19,413	0.83	19,500	0.63	81
1	90,000	0.82	19,129	0.83	19,500	0.56	77
2	90,000	0.86	18,444	0.87	19,500	0.77	77
3	90,000	0.91	17,773	0.91	19,500	0.85	77
Overall	450,000	0.91	94,479	0.91	97,500	0.82	82

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

Split-Half Reliability

In classical test theory, before the advent of digital computers automated the calculation of internal consistency reliability measures such as Cronbach's alpha, approximations such as the split-half method were sometimes used. A split-half reliability coefficient is calculated in three steps. First, the test is divided into two halves, and scores are calculated for each half. Second, the correlation between the two resulting sets of scores is calculated; this correlation is an estimate of the reliability of a half-length test. Third, the resulting reliability value is adjusted, using the Spearman-Brown formula,[1] to estimate the reliability of the full-length test.

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

Table 13 and Table 14 contain split-half reliability estimates for Star Early Literacy, calculated from the 2015–2016 school year data. Split-half scores were based on the odd- and the even-numbered items were calculated. Adjusting the half-length assessments (13 items) to full length assessments (27 items) using the Spearman Brown formula, yielded the split-half reliability estimates displayed in Table 13 and Table 14. As shown in both tables, the split-half reliability estimates are equally as high in both the Unified and the Star Early Literacy scales. Moreover, they are also very similar to the generic reliabilities providing further proof of the high reliability of Star Early Literacy scores.

Retest (Alternate Forms) Reliability

Another method of evaluating the reliability of a test is to administer the test twice to the same examinees. Next, a reliability coefficient is obtained by calculating the correlation between the two sets of test scores. This is called a retest, or "alternate forms," reliability coefficient. This approach was used for Star Early Literacy using data from the 2015–2016 school year.

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 as evident in both Table 13 and Table 14. Irrespective, the test-retest reliabilities are still high.

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

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 30 means that if a student were tested repeatedly, his or her scores would fluctuate within 30 points of his or her first score about 68 percent of the time, and within 60 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 Early Literacy tests differ from traditional tests in at least two respects with regard to the standard error of measurement. First, Star Early Literacy 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 Early Literacy 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 Early Literacy 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.

Two different sources of data are available for estimating the aggregate standard error of measurement of Star Early Literacy Scaled Scores:

1. The averages and standard deviations of the conditional SEM (CSEM) values calculated by the Star Early Literacy software.
2. Estimates of the global standard error of measurement computed from the estimated generic reliability and the observed standard deviations of the Scaled Scores.

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

   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 15 and Table 16 present two different sets of estimates of the Star Early Literacy measurement error: average conditional standard errors of measurement (CSEM) and global standard error of measurement. One point should be noted here: SEMs calculated using the conditional SEM method probably understate measurement error, since they are based on IRT models that do not fit the response data perfectly and assume that the IRT item difficulty parameters are estimated without error.

All in all, the SEM calculated from the generic reliability coefficients and using the standard deviation of the observed Scaled Scores may be the best estimate of the typical SEM for a single test administration. Table 15 and Table 16 provide estimates of the two SEM calculations for each grade and overall for all grades, Pre-K to Grade 3. The overall average conditional standard error of measurement CSEM) is 31 on the Unified scale (Table 15) and 42 on the Star Early Literacy scale (Table 16). The differences in the CSEMs from the Unified to the Star Early Literacy scale can be attributed to the non-linear transformation in the latter scale. 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 Star Early Literacy Scaled Score nonlinear transformation.

Table 15: Star Early Literacy Standard Errors of Measurement on the Unified Scale from 2015–2016 School Year Data

		Standard Error of Measurement, Unified Scale		
		Conditional		
Grade	Sample Size	Average	Standard Deviation	Global
Pre-K	90,000	29	3.7	27
K	90,000	31	4.8	30
1	90,000	32	6.3	31
2	90,000	31	9.6	30
3	90,000	32	13.0	30
Overall	450,000	31	8.3	30

Table 16: Standard Errors of Measurement on the Star Early Literacy Scale from 2015–2016 School Year Data

Grade	Sample Size	Standard Error of Measurement, Star Early Literacy Scale		
		Conditional		Global
		Average	Standard Deviation	
Pre-K	90,000	43	13.9	45
K	90,000	49	10.5	48
1	90,000	45	11.6	44
2	90,000	38	14.0	39
3	90,000	36	15.3	37
Overall	450,000	42	14.0	44

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 test is valid to the extent that there are evidentiary data to support claims as to what the test measures; 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, or is suitable for its intended uses. Star Early Literacy was designed to measure a wide range of skills that culminate in the ability to read in English. A first step in building the case for its validity has to do with the content of the test items that make up its item bank, and are used in each individual test. As described in "Content and Item Development" on page 10, the original 2,929 Star Early Literacy test items were designed explicitly to consist of indicators of seven specific literacy domains and 41 sets of subordinate skills that comprise them. More recently, Star Early Literacy was substantially revised. The structure of the current version organizes its items into three broad domains that subsume 10 subdomains and 41 component skill sets. The organization of the current version was done after extensive review of multiple sets of content and curriculum standards, including the Common Core State Standards that at one point had been adopted by no less than 45 states within the US.

The content of the item bank and the content balancing specifications that govern the administration of each test together form the basis for Star Early Literacy's "content validity."

This section deals with other evidence of Star Early Literacy's validity as an assessment of early literacy skills. All of the evidence presented here has to do with the relationship of Star Early Literacy scores to external variables that are related to the development of literacy skills. Some of the features that a valid literacy skills assessment should have are listed below.

Scores on the assessment should:

- Increase directly with test-takers' ages
- Increase with grade in school
- Correlate with scores on related assessments, such as:
  - Other tests of readiness and early literacy
  - Early-grade reading tests
  - Teachers' ratings of students' mastery of literacy skills

This section consists of evidence, accumulated to date, of the relationships of Star Early Literacy scores to the kinds of external variables cited above. Such evidence is cumulative, having been collected or observed on the original 7-domain version of Star Early Literacy, and on the current 3-domain/10-subdomain version of the assessment.

Relationship of Star Early Literacy Scores to Age and School Grade

The fundamental literacy skills that Star Early Literacy was designed to measure improve as children mature and as they benefit from instruction. Consequently, if Star Early Literacy is indeed measuring literacy skills along a developmental continuum, Star Early Literacy test scores should increase with age and with years of schooling. Evidence of this relationship was obtained in research preceding the 2001 release of the initial version of Star Early Literacy. This research was replicated in 2012 for the current version of the assessment.

Evidence that scores on the current version increase with student age and grade is presented below in Table 17 and Table 18.

Table 17 presents data on the relationship of scaled scores on the current version to student age, for students less than 10 years old. Table 18 presents similar data related to student grade levels, for the Star Early Literacy design grade range of K–3. As expected, average scores increase with each year of age, and with each grade.

Table 17: Summary Statistics on Scaled Scores by Age Group

Age Group	N	Mean	Standard Deviation
5	103	635	111
6	1,190	692	103
7	1,457	744	102
8	902	799	91
9	422	826	73

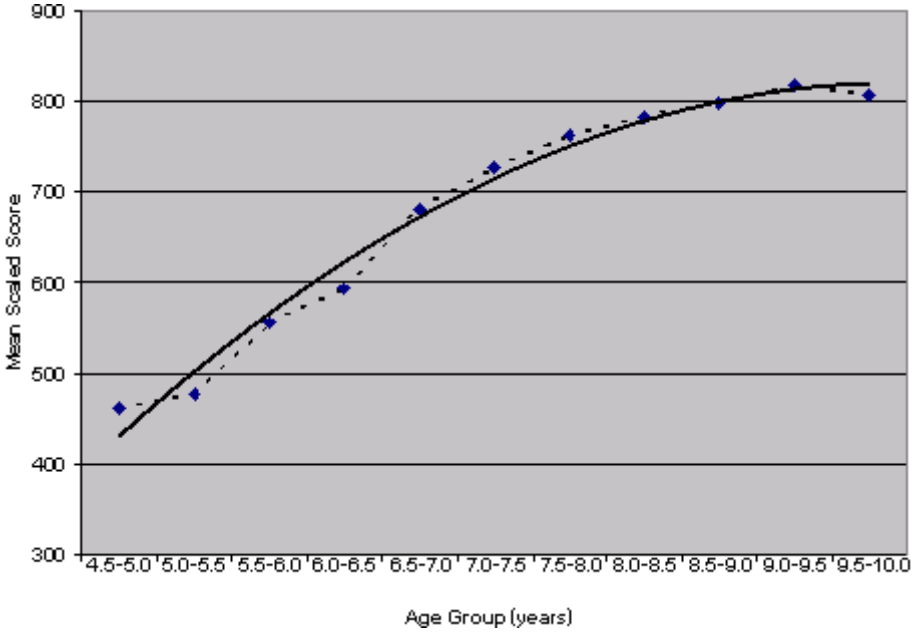
Table 18: Summary Statistics on Scaled Scores by Student Grade

Grade	N	Mean	Standard Deviation
K	2,250	686	101
1	2,782	748	98
2	1,398	818	69
3	946	836	63

As displayed in these tables, adaptive test scores from the 2012 study increased systematically from kindergarten through grade 3. The standard deviation statistics show that score variability was similar from kindergarten through grade 1, but less variable in grades 2 and 3.

Figure 2 displays average Star Early Literacy Scaled Scores for 11 age groups of students participating in the2001 validation of the original version. Each age group spans a six-month age range; the youngest group ranged from 4.5 to 5.0 years old on the date of the Star Early Literacy assessment; the oldest group ranged from 9.5 to 10 years old. As the figure shows, on average, Star Early Literacy Scaled Scores increased directly with students' ages from below 5 to above 9 years old. A small decrease occurred for the oldest age group; this decrease probably reflects the performance of a disproportionate number of low-ability students in the oldest age group.

Figure 2: Star Early Literacy Scaled Scores as a Function of Age; Mean Scaled Scores for 11 Age Groups



Relationship of Star Early Literacy Scores to Skills Ratings

The relationship of scores on the current version of Star Early Literacy to ratings of students' skills was evaluated in the 2012 study. In order to have an independent common measure of literacy skills, Renaissance Learning constructed a 10-item checklist for teachers to use to rate their students on a wide range of competencies related to developing literacy skills. The intent of this checklist was to provide teachers with a single, brief instrument

they could use to rate any student from pre-kindergarten through third grade. In this section, we present the new skills rating instrument itself, its psychometric properties as observed in the study, and the relationship between student skills ratings on the instrument and their Star Early Literacy (SEL) scaled scores.

The Rating Instrument

To gather ratings of literacy skills from teachers, a short list of dichotomous items that represent a hierarchy of skills aligned to the CCSS was constructed. A list of ten skill-related items was assembled. Each participating teacher was asked to rate his or her students on each skill. The rating task was administered by means of an interactive spreadsheet that automatically recorded teachers ratings of each student. The teacher had simply to mark, for each student, any task he/she believed the student could perform. A list of the skills teachers rated for their students is included below.

Skills Rating Items Used in the Spring 2012 Star Early Literacy Skills Ratings Survey

1. Point to each of the words in a 3-word sentence. (CCSS Pre-K)
2. Say “yes” if the words have the last same sounds (rhyme): mop/top (y) down, boy (n) (CCSS Pre-K)
3. Say the letters in your name. (IES Birth-Age 5)
4. Identify the lowercase letter “d.” (CCSS K)
5. Say the sound that begins these words: milk, mouth, mother (/m/) (CCSS-K)
6. Read aloud the printed word “said.” (CCSS Grade 1)
7. Write and spell correctly the word “fish.” (CCSS Grade 1)
8. Read words containing short vowel sounds *bit, tap, hop* and long vowel sounds *bite, tape, hope* (CCSS Grade 1)
9. Read aloud and distinguish the meanings of the printed words “two” and “too.” (CCSS Grade 2)
10. Read on-level text with purpose and understanding (CCSS Grade 2)

The rating worksheet was scored for each student by assigning one point for each performance task marked by the teacher. The range of possible scores was 0 to 10. Teachers completed skills ratings for 6,708 students. Table 19 lists data about the psychometric properties of the ten item rating scale,

overall and by grade, including the correlations between skills ratings and SEL Scaled Scores. The internal consistency reliability of the skills ratings was 0.84, as estimated by coefficient alpha.

Relationship of Scaled Scores to Skills Ratings

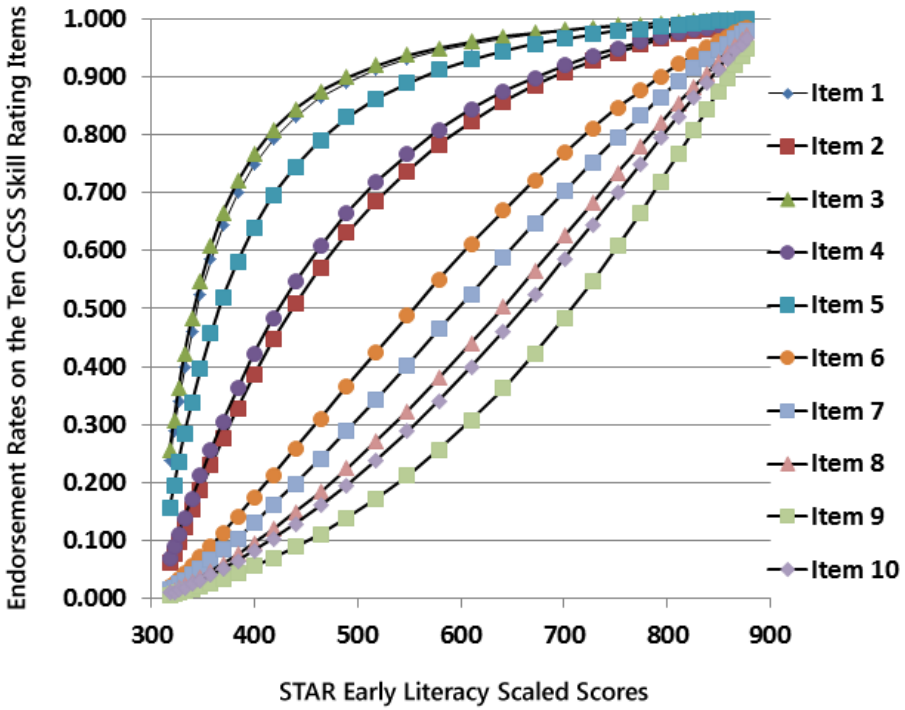
As the data in Table 19 show, the mean skills rating increased directly with grade, from 7.48 at kindergarten to 9.48 at grade 3. Thus, teachers rated kindergarten students as possessing fewer than eight of the ten skills on average. In contrast, the average third grader was rated as possessing almost all of the ten skills. The correlation between the skills ratings and SEL Scaled Scores was significant at every grade level. The overall correlation was 0.58, indicating a substantial degree of relationship between the Star Early Literacy test and teachers' ratings of their students' CCSS literacy skills. Within-grade correlations ranged from 0.46 to 0.60.

Table 19: Teachers' Ratings of Their Students' Common Core State Standards-Aligned Skills, by Grade

Grade	N	Ratings		Scaled Scores		Correlation of Skills Ratings with Scaled Scores
		Mean	Standard Deviation	Mean	Standard Deviation	
K	2,039	7.5	2.4	687	99	0.47
1	2,552	8.7	2.0	749	98	0.58
2	1,243	9.0	1.9	818	69	0.46
3	874	9.5	1.3	839	59	0.60
All	6,708					0.58

Figure 3 displays the relationships of each of the ten rating scale items to SEL Scaled Scores. These relationships were obtained by fitting mathematical models to the response data for each of the ten rating items. Each of the curves in the figure is a graphical depiction of the respective model. As the curves show, the proportion of students rated as possessing each of the ten rated skills increases with Scaled Score.

Figure 3: Relationship of the Endorsement Rates on the Ten CCSS Skill Rating Items to Star Early Literacy Scaled Scores



Relationship of Star Early Literacy Scores to Other Tests

Besides showing the appropriate relationships with age, grade level, and skills ratings by teachers, if Star Early Literacy is indeed measuring literacy skills, its scores should correlate positively with measures of reading, literacy, and readiness. To evaluate this, standardized reading and other test scores were collected for students participating in the validation study of the original version, in 2001. In that study, scores recorded on a variety of standardized reading tests were entered by teachers, using a special worksheet provided for this purpose. Subsequent to the 2001 study, additional data have been collected, on both the original and current versions, that show the relationship of Star Early Literacy scores to scores on other tests. Below, results from the original validation study are shown, followed by results from recent studies.

Validation Study Data

As part of the original Validation Study, participating teachers were asked to provide students' scores from a variety of other tests. Renaissance Learning provided the teachers with a special worksheet to record such scores. In

addition to reading test scores, scores on a number of other tests were obtained for many of the students participating in the Validation Study. These tests included other measures of early literacy as well as tests of readiness, social skills, and other attributes.

Usable scores were received for over 2,400 students on 20 different test series administered in the fall or spring of the 2000 school year or the spring of 2001. Most of the reported scores were either NCE scores or Scaled Scores. From the usable data, 61 correlations with Star Early Literacy were computed. The number of correlations ranged from 10 at the kindergarten level to 22 at grade 3. No external test scores were available for pre-kindergarten students.

As part of the ongoing efforts to provide evidence for the validity of Star Early Literacy scores, further research studies have been carried out. Additional concurrent validity studies have been undertaken, and the results were added to the overall results. Concurrent validity was operationally defined as the extent to which Star Early Literacy scores correlated with scores on external measures, when both tests were given within the same two-month period.

In addition, predictive validity studies have been undertaken to provide some measure of the utility of using Star Early Literacy for predicting later outcomes. Predictive validity was defined as the extent to which scores on the Star tests predict 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. It provided an estimate of the linear relationship between Star scores and scores on measures 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.

Detailed results of the concurrent and predictive correlational analyses are documented in "Appendix: Detailed Validity Data" on page 108. This chapter presents summaries of key data.

Table 20 on page 56 presents summaries of concurrent and predictive correlation coefficients between the scores on Star Early Literacy and each of the other test instruments (external measures) for which data were received in the course of the 2001 validation study. Data were collected for a wide range of external measures.

The averages of the concurrent validity correlations within grade were 0.64, 0.68, 0.52, and 0.57 for grades K–3 respectively. The overall concurrent correlation was 0.59. Criterion measures included the following:

Criterion Measures for Concurrent and Predictive Validity Analyses
Brigance K & 1 Screen for Kindergarten and First Grade Children
Canadian Achievement Test
Child Observation Record (OCR)
Developing Skills Checklist (DSC)
Developmental Indicators for the Assessment of Learning (DIAL)
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)
Florida Comprehensive Assessment Test (FCAT)
Gates-MacGinitie Reading Test (GMRT)
GRADE
Iowa Test of Basic Skills (ITBS)
Metropolitan Early Childhood Assessment Program (MKIDS)
Michigan Literacy Progress Profile (MLPP)
Running Record (a Reading Recovery component)
Stanford Achievement Test
STAR Reading
TerraNova
Test of Phonological Awareness (TOPA)
Texas Primary Reading Inventory (TPRI)

The averages of predictive validity coefficients for pre-K–3 were, respectively, 0.57, 0.52, 0.62, 0.67, and 0.77. The overall average predictive validity coefficient across the grades was 0.58.

Table 20: Summary of Concurrent and Predictive Validity: Star Early Literacy Correlations with Tests Administered in Spring 2001

Concurrent Correlations with Other Tests, Grades K–3					
Grades	**All**	**K**	**1**	**2**	**3**
Number of students	1,376	198	281	513	384
Number of coefficients	34	6	7	12	9
Average validity	–	0.64	0.68	0.52	0.57
Overall average	0.59				
Predictive Correlations, K–3					
Grades	**All**	**K**	**1**	**2**	**3**
Number of students	61,443	557	30,423	5,370	558
Number of coefficients	51	4	15	15	2
Average validity	–	0.57	0.52	0.67	0.77
Overall average	0.58				

Meta-Analyses of the Validation Study Validity Data

Meta-analysis is a set of statistical procedures that combines 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, as well as standard errors and confidence intervals, both overall and within grades. To conduct a meta-analysis of the Star Early Literacy validation study data, the 85 correlations observed in the Star Early Literacy 2001 validation study and documented in previous editions of the technical manual were analyzed using a fixed effects model and rescaled using the Fisher r-z transformation. The results are displayed in Table 21. The table lists results for the correlations within each grade, as well as results with all four grades' data combined.

For each set of results, the table lists an estimate of the true validity, a standard error, and the lower and upper limits of a 95 percent confidence interval for the validity coefficient.

Using the validation study data, the overall estimate of the validity of Star Early Literacy is 0.60, with a standard error of 0.02. The true validity is

estimated to lie within the range of 0.57 to 0.62, with a 95 percent confidence level. The probability of observing the 85 correlations reported in Table 20, if the true validity were zero, is virtually zero. Because the 85 correlations were obtained with widely different tests, and among students from four different grades, these results provide support for the validity of Star Early Literacy as a measure of early reading skills.

Table 21: Results of the Meta-Analysis of Star Early Literacy Correlations with Other Tests from the Validation Study

Grade	Effect Size		95% Confidence Level	
	Validity Estimate	Standard Error	Lower Limit	Upper Limit
Kindergarten	0.56	0.06	0.50	0.66
Grade 1	0.64	0.05	0.58	0.69
Grade 2	0.57	0.04	0.52	0.62
Grade 3	0.60	0.03	0.55	0.64
All Grades	0.60	0.02	0.57	0.62

Evidence of Technical Adequacy 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 Early Literacy is widely used for both purposes. This chapter includes technical information about Star Early Literacy'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 Early Literacy 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 Early Literacy has a "convincing" level of evidence (the highest rating awarded) or "partially convincing" (the next highest) in the core screening and psychometric categories, including classification accuracy, reliability, and validity. CRTI also notes that the extent of the technical evidence is "Broad" (the highest rating awarded) and notes that not only is the overall evidence compelling, but there are disaggregated data as well that shows Star Early Literacy works equally well among subgroups. The most recent reviews by NCII indicate that there is fully "convincing" evidence of Star Early Literacy'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 Early Literacy 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 Early Literacy 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 Early Literacy 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."[1]

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 Early Literacy 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[2] are generally consistent in how

1. https://rti4success.org/essential-components-rti/universal-screening
2. http://www.rtinetwork.org/learn/research/universal-screening-within-a-rti-model

measurement tools should be evaluated for their appropriateness as screeners. Key categories include the following:

1. **Validity and reliability.** See the chapter on Reliability and Measurement Precision, other sections of this Validity chapter, and "Appendix: Detailed Validity Data" for evidence of Star Early Literacy reliability and validity.
2. **Practicality and efficiency.** Screening measures should not require much teacher or student time. Because most students can complete a Star Early Literacy test in 10–15 minutes or less, and because it is group administered and scored automatically, Star Early Literacy 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?

One of the key purposes of Star Early Literacy is to ensure that students are on a pathway to become independent readers, so an important test is whether it can predict later reading difficulties. To evaluate Star Early Literacy in this context, a study was conducted to determine the extent to which it could accurately classify students using Star Reading benchmarks. The sample included 2,252 students who took Star Early Literacy in at the end of Kindergarten, grade 1, 2, or 3, and then took Star Reading one year later. A linking study was conducted to establish cut points on the Star Early Literacy scale that corresponded to the criterion, which was the 25th percentile on Star Reading, a commonly used threshold indicating a strong probability of reading difficulties. The results indicated a reasonably strong level of accuracy, considering that predictions were made across school years. Sensitivity and Specificity were 73% and 74%, respectively, and the Area Under the ROC Curve indicated an overall level of accuracy to be between 0.80 and 0.87 by grade.

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."[3]

3. https://intensiveintervention.org/ncii-glossary-terms#Progress Monitoring

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 been 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"[4] as recommended by Shapiro (2008), and Star Early Literacy 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. A summary of those characteristics and relevant information about Star Early Literacy is provided below.

1. **Evidence of psychometric quality.**

    a. **Reliability and validity.** Star Early Literacy's reliability and validity evidence is presented in this manual's chapter on Reliability, other sections of this Validity chapter, and Appendix: Detailed Validity Data.

    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 2008 by Renaissance Learning to evaluate reliability of slope for at-risk students who were being progress monitored during the 2007–08 school year. Specifically, the sample included 23,254 students who 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 Early Literacy 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

4. 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.

least squares regression. Then, the odd and even slopes were correlated. Table 22 summarizes the Pearson correlation coefficients by grade, indicating a consistently strong association between even and odd numbered test slopes.

Table 22: Star Early Literacy Reliability of the Slope Coefficients by grade, K–3

Grade	n	Coefficient
K	11,206	0.75
1	9,577	0.73
2	1,938	0.71
3	533	0.75

2. **Produce a sufficient number of forms.** Because Star Early Literacy 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 Early Literacy 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.

   Overall and grade-level results are summarized in the Reliability chapter.

3. **Practicality and efficiency.** As mentioned above, most students complete Star Early Literacy in 10–15 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 Early Literacy 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. **Percentile Rank and Scaled Score.** Educators may also enter custom goals using Percentile Rank or Scaled Score metrics. Default Percentile Rank benchmarks are set to the 10th percentile (indicating severely at-risk, and a candidate for urgent intervention), 25th percentile (indicating at-risk, and a candidate for intervention), and the 40th percentile. Between the 26th and 40th percentiles, students are labeled "on watch," and above the 40th percentile students are considered to be working approximately at grade level, or above grade level. While there are no universally defined rules for setting normative benchmarks, these benchmarks were set based on consultation with experts in school psychology and are generally consistent with benchmarks used by other assessments for RTI/MTSS purposes.

Additional research on Star Early Literacy as a progress monitoring tool. A study by Cormier & Bulut (manuscript under review) evaluated Star Reading as a progress monitoring tool, the first known study to evaluate and recommend empirically-based administration rules to inform instructional decisions for a computer adaptive assessment. The methodology used in that study was applied by Renaissance staff to Star Early Literacy data. The sample included 1.4 million students in grades K–3 who took Star Early Literacy in the 2014–15 or 2015–16 school years. The study concluded the following:

- 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 Early Literacy for progress monitoring purposes.

- Sufficiently reliable progress monitoring slopes could be generated in as few as five Star Early Literacy administrations.
- The duration of Star Early Literacy 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. Decisions can be made after at least five administrations as early as 10 weeks (Kindergarten) and up to 20 weeks (Grade 3).
- 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.

Ad Hoc Studies of Star Early Literacy Validity

Subsequent to publication of Star Early Literacy in 2001, additional external validity data have become available, both from users of the assessment, and from special studies conducted by Renaissance Learning. This section provides summaries of some of those new data, along with tables of results. Data from three sources are presented here: These were studies of the relationship between Star Early Literacy and 1) Running Record scores, 2) Michigan Literacy Progress Profile (MLPP) scores, and 3) DIBELS, GRADE, and TPRI scores.

Running Record

Running Records are a systematic notation system for teacher observations of children's reading of new text. Use of the Running Record is one component of Reading Recovery, a program pioneered by Marie Clay in New Zealand and now widely used in the US and elsewhere. In early 2002, kindergarten and first grade teachers in a Michigan elementary school administered Star Early Literacy to 72 students who had recently been assessed using the Running Record. The relationship between the two sets of test scores is strong and clear: Star Early Literacy scores varied directly with children's reading proficiency as measured by the Running Record.

As Star Early Literacy scores increased, Running Record scores increased as well. The Pearson correlation between them in this student sample was 0.72.

Michigan Literacy Progress Profile (MLPP)

Developed by the Michigan Department of Education, MLPP is a comprehensive assessment system for preschool to third-grade students. The MLPP is intended to be administered multiple times (3–4) per school year and to provide teachers with a picture of individual students' literacy so that they can target instruction and help each student develop.

Two Michigan elementary schools that used both MLPP and Star Early Literacy provided data for a study from Fall 2003 (n = 245) and Spring 2004 (n = 219). Because the MLPP consists of several individual tests and has no overall score, the correlation between the two assessment systems had to be conducted test by test. The results, in Table 23 below, revealed statistically significant and generally high correlations at significant levels in both Fall 2003 and Spring 2004.For tests given in Fall 2003 the correlation coefficients between the MLPP tests and Star Early Literacy Scaled Scores were between 0.56 and 0.82. The strength of this relationship indicates that Star Early Literacy measures a significant number of the same skills measured by the MLPP tests. For Spring 2004 scores, the range of correlations is similar. Most of the Spring 2004 correlations have been corrected for range restriction, to account for ceiling effects on the MLPP tests.

A complete description of the MLPP Validation Study, including correlation of MLPP tests with Star Early Literacy Domain Scores, is presented in *Correlation Between Michigan Literacy Progress Profile and Star Early Literacy,* an April 2005 report.

Table 23: Star Early Literacy Scaled Score Correlations with Michigan Literacy Progress Profile Raw Scores, Combined K and 1

	Fall 2003		Spring 2004	
Test	**N**	**r**	**N**	**r**
Concepts of Print	245	0.74	219	0.74
Letter Name	245	0.76	219	0.72
Letter Sounds	245	0.80	219	0.74
Word List	245	0.62	219	0.81
Known Words	245	0.70	219	0.66
Rhyming	245	0.56	219	0.53
Dictation	245	0.82	219	0.76
Segmenting	245	0.69	219	0.57
Blending	245	0.71	219	0.73

DIBELS, GRADE, and TPRI

In September and October, 2004, Renaissance Learning conducted a study of the relationships of Star Early Literacy scores and scores on three widely used early literacy assessments: DIBELS,[5] TPRI,[6] and GRADE.[7] These assessments were chosen for study because they measure most or all of the five critical skills identified in the 2000 report of the National Reading Panel: Phonological Awareness, Phonics, Vocabulary, Text Comprehension, and Fluency. Two of them, DIBELS and TPRI, are widely used for assessment within Reading First programs.

Following is a short summary of the tests administered within each assessment. Eight schools from six different states participated in this study, administering DIBELS, GRADE, Star Early Literacy and TPRI to students in kindergarten and first and second grade. Approximately 200 students were tested at each grade. Correlations of selected scores with Star Early Literacy scaled scores are reported in Table 24.[8] As the data in Table 24 show, in the kindergarten sample, Star Early Literacy correlated highest with Letter Naming in DIBELS; with Phonological Awareness, Listening Comprehension, and Phoneme-Grapheme Correspondence in the GRADE test; and with Blending Word Parts and Blending Phonemes in TPRI.

At first grade, the correlations were moderate to high with all DIBELS tests except Phoneme Segmentation, with all of the GRADE tests administered, and with all TPRI tests except Letter Sounds and Comprehension. Correlations with oral reading fluency were among the highest at first grade, despite the fact that Star Early Literacy does not include an oral reading component.

At second grade, all correlations were moderate to high, except the 0.30 correlation with TPRI Comprehension. As in the first grade sample, correlations were high with both of the oral fluency measures (DIBELS and TPRI). The low correlation with TPRI Comprehension is contradicted by the correlations with the GRADE Comprehension measure, and with DIBELS Retell Fluency measure, which is characterized as a comprehension measure.

---

5. Dynamic Indicators of Basic Early Literacy Skills (University of Oregon, Institute for Development of Educational Achievement, 2002).
6. Texas Primary Reading Inventory 2004–2006 (Texas Education Agency and the University of Texas System, 2003).
7. Group Reading Assessment and Diagnostic Evaluation (American Guidance Service, Inc., 2001).
8. Both teacher discretion in the choice of which DIBELS tests to administer and TPRI test-to-test branching rules resulted in numerous cases of incomplete sets of test scores. To improve the statistical accuracy of some correlations, missing scores were imputed. Correlations reported here were calculated in the imputed data sets.

Table 24: Star Early Literacy Correlations with DIBELS, GRADE, and TPRI

	Kindergarten		Grade 1		Grade 2	
	r	n	r	n	r	n
DIBELS						
Initial Sounds	0.24	214	–	–	–	–
Letter Naming	0.45	214	0.58	198	–	–
Phoneme Segmentation	0.30	214	0.29	198	–	–
Nonsense Words	0.36	214	0.71	198	0.59	201
Word Usage	0.36	214	0.55	198	0.44	201
Retell	–	–	0.64	198	0.67	201
Oral Reading	–	–	0.78	198	0.72	201
GRADE						
Phonological Awareness	0.54	214	–	–	–	–
Sound Matching	0.44	214	–	–	–	–
Rhyming	0.53	214	–	–	–	–
Early Literacy Skills	0.34	214	–	–	–	–
Print Awareness	0.35	214	–	–	–	–
Letter Recognition	0.27	214	–	–	–	–
Same and Different Words	0.39	214	–	–	–	–
Phoneme-Grapheme Correspondence	0.44	214	–	–	–	–
Vocabulary	–	–	0.73	198	0.69	201
Word Meaning	–	–	0.71	198	0.61	201
Word Reading	0.35	214	0.67	198	0.64	201
Comprehension	–	–	0.68	198	0.76	201
Sentence Comprehension	–	–	0.63	198	0.72	201
Passage Comprehension	–	–	0.65	198	0.70	201
Listening Comprehension	0.45	214	0.50	198	0.52	201

Table 24: Star Early Literacy Correlations with DIBELS, GRADE, and TPRI

	Kindergarten		Grade 1		Grade 2	
	r	n	r	n	r	n
TPRI						
Screening: Graphophonemic Knowledge	0.23	214	–	–	–	–
Screening: Phonemic Awareness	0.33	214	–	–	–	–
Rhyming	0.26	214	–	–	–	–
Blending Word Parts	0.64	214	–	–	–	–
Blending Phonemes	0.56	214	–	–	–	–
Detecting Initial Sounds	0.39	214	–	–	–	–
Detecting Final Sounds	–0.14	214	–	–	–	–
Letter Name Identification	0.36	214	–	–	–	–
Phonemic Awareness	0.35	214	–	–	–	–
Listening Comprehension	0.34	214	–	–	–	–
Letter Sounds	0.16	214	0.34	198	–	–
Word Reading	–	–	0.69	198	0.53	201
Graphophonemic Knowledge	0.23	214	–	–	0.64	201
Story Number	0.03	214	0.69	198	0.50	201
Fluency	–	–	0.70	198	0.67	201
Comprehension	–	–	0.32	198	0.30	201

Predicting Reading Performance from Star Early Literacy Scores

Betts and McBride (2006) evaluated Star Early Literacy's validity for predicting future outcomes, including scores on later measurements of early literacy skills as well as performance on a measure of reading achievement. Theirs was a longitudinal study, in which six age cohorts of school children were followed for two years. The age cohorts included children in three initial-year school grades: kindergarten, grade 1, and grade 2. Students in each cohort took Star Early Literacy on multiple occasions each year, to monitor the development of their early literacy skills, and took Star Reading in the final year to measure their reading achievement.

This study evaluated developmental validity, as well as the predictive validity of Star Early Literacy with respect to later reading ability. Predictive validity was assessed in two ways: first, with respect to later scores on the same measure across a single school year; second, with respect to scores on Star

Reading taken two years after the initial assessment of early reading skills. This provided estimates of predictive validity across three time points during the kindergarten, first, and second grade school years of early reading skills.

The cohorts' test records were compiled from a large database of over 40,000 users that spanned school years 2001–2002 through 2004–2005. The student records used for this study were from 130 schools representing 30 different states and included urban, rural and suburban school districts.

Table 25 displays correlations among Star Early Literacy test scores at different occasions, and of Star Early Literacy test scores with Star Reading scores. Star Early Literacy–Star Reading correlations corrected for measurement error (Crocker & Algina, 1986) are also included.

Star Early Literacy scores taken during the fall measurement point at the beginning of the school year are significantly predictive of Star Early Literacy scores at both the winter and spring measurement occasions. Similarly, the winter assessment was significantly predictive of the spring assessment. This indicates that early literacy scores within a school year are highly predictive of later scores. In addition, the Star Early Literacy scores at each occasion were moderately to highly related to reading scores two years after the original assessment occasion; overall, Star Early Literacy scores from any season correlated between 0.71 and 0.79 with Star Reading tests taken two years later. These results are consistent for each of the subgroups partitioned by initial grade level.

In summary, the data displayed in Table 25 provide support from a substantial longitudinal study, for both the validity of Star Early Literacy as a measure of developing skills, and for its long-term validity for predicting later reading achievement.

Table 25: Validity Coefficients of Star Early Literacy with Itself at Later Time Points and with Star Reading over a Two-Year Period for all Cohort Groups Combined and Separately

Grade Groups		Star Early Literacy Initial-Year Scale Scores			Star Reading End-Year Scale Scores[a]
		Fall	Winter	Spring	Fall
Total Sample	Mean	593.07	641.39	694.41	318.47
	Std Dev[b]	128.51	120.03	115.76	165.83
	N	2,730	2,978	3,384	4,028
	ρxx	0.87	0.85	0.86	0.93

Table 25: Validity Coefficients of Star Early Literacy with Itself at Later Time Points and with Star Reading over a Two-Year Period for all Cohort Groups Combined and Separately

Grade Groups		Star Early Literacy Initial-Year Scale Scores			Star Reading End-Year Scale Scores[a]
		Fall	Winter	Spring	Fall
K-2[c]	Mean	490.87	555.23	615.13	218.27
	Std Dev	86.28	92.35	96.26	127.45
	N	1,024	1,230	1,501	1,312
	ρxx	0.75	0.71	0.72	0.92
1-3[d]	Mean	613.70	684.08	745.27	340.25
	Std Dev	97.05	93.54	88.54	149.92
	N	1,082	1,322	1,359	1,749
	ρxx	0.72	0.74	0.79	0.90
2-4[e]	Mean	725.03	757.68	789.64	415.01
	Std Dev	91.80	91.37	76.79	167.64
	N	624	426	524	967
	ρxx	0.83	0.78	0.83	0.89

a. All coefficients are statistically significant ($p < 0.001$).
b. Star Early Literacy scores were taken within the same school year.
c. Star Reading was taken two years after the fall administration of Star Early Literacy.
d. Corrected for measurement error.
e. "r" indicates the validity coefficient rounded to two decimal places.

Concurrent Validity of Estimated Oral Reading Score

During the fall of 2007 and winter of 2008, 25 schools across the United States that were using both Star Early Literacy and DIBELS Oral Reading Fluency (DORF) participated in research to provide evidence supporting the validity of Star Early Literacy's Estimated Oral Reading Fluency (Est. ORF) score. The schools were asked to ensure that students were tested on both Star Early Literacy and DORF within a 2-week time interval during September and January.

The analysis was undertaken to estimate the extent to which the Est. ORF scores on Star Early Literacy accurately predicted the observed DIBELS Oral Reading Fluency scores. Both the Est. ORF score on Star Early Literacy and DORF provide estimates of the students' oral reading fluency expressed as the number of words read correctly within one minute (WCPM) on a grade-level-appropriate connected text passage. The Est. ORF score is an estimate

based on the student's performance on Star Early Literacy, while the DORF score is a direct measure from a set of standardized grade-level passages.

Analysis was done on each grade independently because DORF passages are assigned to specific grade levels and therefore are not interpretable across grades.

The 25 schools in the sample came from nine 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 = 3,221). The distribution of students by grade was as follows: 1st grade 2,028, 2nd grade 729, and 3rd grade 464.

Correlations between the Est. ORF and DORF WCPM are displayed inTable 26 along with the mean difference, standard deviation of differences, and the sample sizes. Correlations were moderately high, ranging from 0.64 to 0.69. Mean differences between Est. ORF and WCPM ranged from 3.99 to −7.29, indicating that at grade 1 the Est. ORF tended to over-estimate the students' reading fluency by about 4 words per minute, whereas it tended to underestimate WCPM in grade 3 by about 7 words. These differences were small but statistically significant (all $p < 0.001$).

Table 26 displays results of a subsequent study of Grade 1 students. That study administered the Aims Web Oral Reading Fluency test. The correlation between Est. ORF and the Aims Web oral reading fluency measure (WCPM) was 0.77; the mean difference between the Star estimate and the observed fluency rate was 1.2 words.

Table 26: Relations between Star Early Literacy Est. ORF and DIBELS Oral Reading Fluency

a. 25-schools original study				
		Differences (Est. ORF Minus DORF)		
Grade	Correlation	Average	Standard Deviation	Sample Size
1	0.69	4.0	21	2,028
2	0.64	−3.7	23	729
3	0.64	−7.3	28	464
b. Single school follow-up study				
		Differences (Est. ORF Minus Aims Web ORF)		
Grade	Correlation	Average	Standard Deviation	Sample Size
1	0.77	1.2	21	117

Equivalence and Validity of the Current and Original Star Early Literacy Versions

The principal evidence related to the equivalence of the current and previous version consisted of score distributions and correlations. Two kinds of Star Early Literacy scores were examined:

- **Rasch ability estimates.** These are the fundamental scores on the adaptive Star Early Literacy tests. They are recorded as decimal-valued real numbers; typical values range from –6.00 to +6.00.
- **Star Early Literacy Scaled Scores.** These are the principal scores reported for Star Early Literacy tests. Scaled Scores are non-linear but monotonic transformations of the Rasch scores; they take values from 300 to 900.

Table 27 lists summary statistics on the scores from current and original versions—including Rasch ability and Scaled Scores—for all grades combined, including grades pre-K through grade 5. The scores for the original version are averages of 1 to 3 scores recorded during the research testing window; students with more than 3 service tests during the research window were excluded from all analyses in this section.

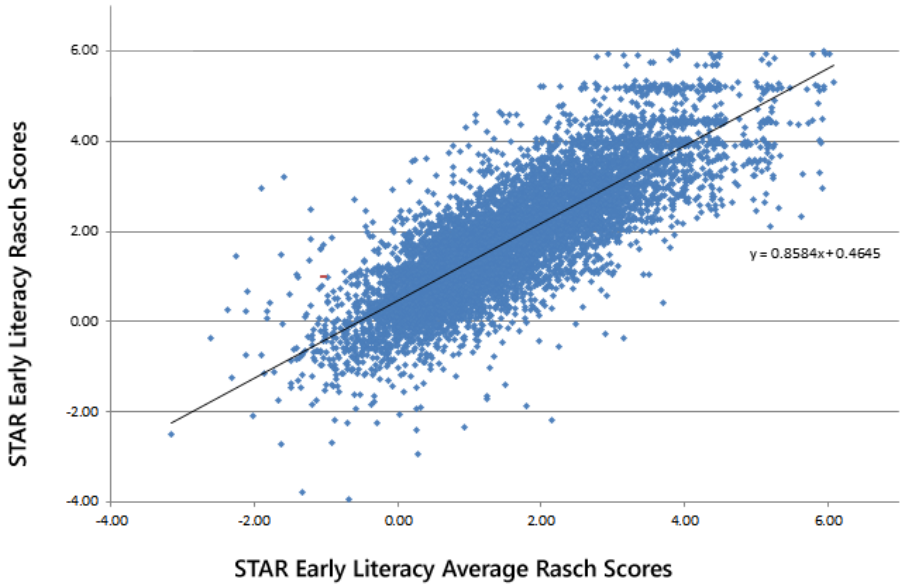
The mean Scaled Score on the new (current) version was 755, 13 points higher than the average Scaled Score on the original version, 742. Similarly, the mean Rasch ability score was 1.88, which was 0.18 logits (Rasch scale units) higher than the original version average of 1.70.

Table 27: Summary Statistics on Original and Current Versions, All Grades Combined

Variable	N	Mean	Standard Deviation
Star Early Literacy Scaled Score	7,070	755	104
Star Early Literacy Rasch Score	7,070	1.88	1.37
Average Old version Scale Score	7,070	742	1.06
Average Old version Rasch Score	7,070	1.70	1.33

Figure 4 displays the scatterplot of the original version Rasch ability scores versus the current version scores. The plot illustrates the strong linear relationship between scores on the two tests.

Figure 4: Plot of Star Early Literacy Original vs. Current Version Rasch Scores



The observed correlation coefficients are attenuated by the unreliability in both tests. Correcting the Rasch and scaled score correlations for attenuation, using the formula developed by Spearman (1904), results in corrected correlations shown in Table 28.

The corrected correlations shown in Table 28 are high, but substantially less than 1.00. This implies that the current and original versions are measuring highly related, but not identical, constructs.

To make scores on the current version comparable to Star Early Literacy original version scores, a scale linkage was performed, using a linear equating approach. The resulting linkage equation is:

Equivalent Star Early Literacy Rasch Ability Score =
0.9746 × Current Version Rasch Score – 0.1350

Table 28: Correlations between Current and Original Version Scaled Scores and Rasch Scores, Corrected for Attenuation

	Average Service Scaled Score	Average Service Rasch Score
Scaled Score	0.91	
Rasch Score		0.93

The Validity of Early Numeracy Test Items as Measures of the Early Literacy Construct

Since its initial version released in 2001, Star Early Literacy's item bank has always included some items that could be characterized as measuring early numeracy skills. However, until the release of the current version in 2012, the test blueprint did not prescribe specific numbers of numeracy items to be administered. The current version's item bank contains more numeracy items than previous versions, and its test blueprint prescribes administering numeracy items to every student, as the final five items of the 27-item test. With this change came a need to evaluate whether the five-item block of Early Numeracy (EN) items measures the same construct as the Early Literacy (EL) items that constitute the first 22 items of the current assessment. Exploratory factor analysis was used to investigate this.

The data for the factor analysis were collected during the Star Early Literacy research study conducted in early 2012. During that study, several thousand students took the current version, and the majority of those students also took the previous version within a few days before or afterwards. The factor analysis was based on data from 6,785 students who took both versions; student records with complete data on both tests were used for the factor analysis.

The current version comprises 27 items administered adaptively. Sequentially, the first 22 items are EL items and last 5 are EN items. For this analysis, the items were split into three groups, as follows: the first group contained the first 17 of the 22 EL items, the second group contained the last 5 of the 22 EL items, and the last group contained the 5 EN items.

For purposes of the exploratory factor analysis, each of these three item groups was scored separately, so there were two EL scores (based on 17 and 5 items, respectively) and one EN score, also based on 5 items. This grouping was motivated by the desire to correlate scores from the 5 EN items with scores from a corresponding number of EL items.[9] In addition, there were test scores on the original version of STAR Early Literacy and skills ratings for all students in the study. The skills ratings were provided by teachers of the participants.

9. The purpose of this was to account for the fact that any correlation based on just 5 items would be lower than a correlation based on 22 items, even if the shorter and the longer item blocks measured identical constructs. Calculating both EL and EN scores, each based on just five items, provides a means of comparing the EL and EN blocks on an equal footing.

The variables that were included in the factor analysis were:

- EL17—scores on the first 17 items of the 22 EL items.
- EL5—scores on the last 5 of the 22 EL items.
- EN5—scores on the 5 EN items.
- Older Version SEL Scores—scores on the previous SEL version.
- Rating—the composite skills rating scores.

All the scores were Rasch ability estimates except the rating composite scores, which were summed scores. Rasch ability estimates were used for the analysis, rather than scale scores, because the test itself employs the Rasch metric for adaptive item selection, as well as scoring. Scale scores are transformations of Rasch estimates to a metric more familiar to teachers, students, and parents.

The correlations between all five scores are shown in Table 13. Although all the correlations are positive, of particular interest is the correlation between the EN5 scores and the EL5 and EL17 scores. There is, for instance, a positive and moderately strong association between the EN5 and the EL5 scores, r = 0.59, indicating they measure the same construct to some extent. The reader should note that this correlation is based on scores computed from only 5 items, a situation that would be expected to attenuate the correlation. The correlation between the EL17 scores and the EN5 scores was 0.64 which was greater than that between the EL5 scores and the EN5 scores (r = 0.59).

Table 29: Score Correlations (N = 6,438)

	EL17	EL5	EN5	Service	Rating
EL17	1				
EL5	0.73*	1			
EN5	0.64*	0.59*	1		
Older Version SEL Scores	0.77*	0.69*	0.61*	1	
Rating	0.50*	0.47*	0.43*	0.52*	1

Note: * indicates significance with p-value < 0.0001.

Often, correlations do not tell the whole story. Next, we present the results of the factor analysis that examined the relationship of the 5 variables to the underlying construct. Factor analysis can identify as many as 4 distinct latent attributes (factors) that account for the intercorrelations among 5 variables. One standard practice is to disregard any factors that have an eigenvalue less than 1.0; using that criterion in this case resulted in retaining only one

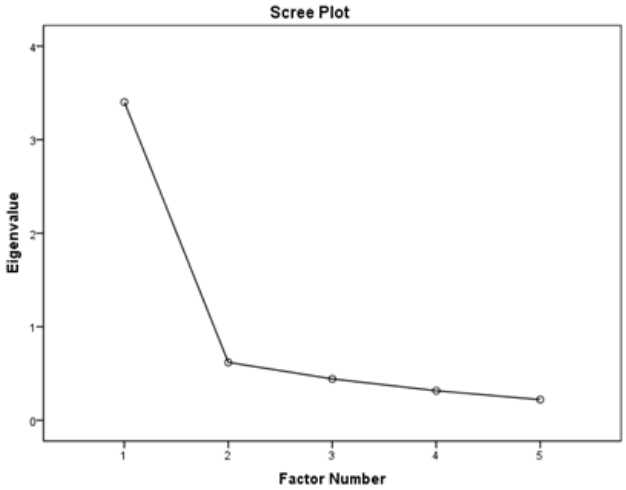
factor. Several variants of factor analysis were applied and they all resulted in one factor. The factor loadings are shown in Table 30. All the scores strongly loaded on the one dominant factor, which is an indication that the EL and the EN items both measure the same underlying construct. The factor represents the construct of interest whereas the loadings represent the correlation between the scores on each variable (e.g., EN5) and the construct.

Table 30: Factor Loadings

Variable	Factor Loading
EL17	0.89
EL5	0.81
EN5	0.72
Older Version SEL Scores	0.86
Rating	0.58

The scree plot of the extracted factors is presented in Figure 5, which provides a visual confirmation that there is only one dominant factor. The empirical evidence strongly suggests one dominant factor representing the construct of interest. The EN items demonstrably measure the same construct as do the original version of SEL and the EL items within the current version. This supports including the EN items in the calculation of SEL scores.

Figure 5: Scree Plot of the Extracted Factors



The Dimensionality of Star Early Literacy

Results of the exploratory factor analyses reported immediately above provide evidence that the five early numeracy items in the current version are part of

the same construct underlying the early literacy items themselves, as well as teachers' ratings of their students on a 10-item early literacy skills checklist. The desire for additional evidence of the dimensionality of the current, 27-item version of SEL led to an analysis of subscores on each of the 10 SEL sub-domains.

The ten sub-domains are fall under the three broad domains: *Word Knowledge and Skills* (the first 7 subdomains below); *Comprehension Strategies and Constructing Meaning* (Sentence-level comprehension and Paragraph- level comprehension); and *Numbers and Operations* (Early Numeracy).

Recall that Star Early literacy items are selected adaptively for each student, from a large bank of early literacy test items, each of which is aligned to one of ten content sub-domains:

- Alphabetic Principle – AP
- Concept of Word – CW
- Visual Discrimination – VS
- Phonemic Awareness – PA
- Phonics – PH
- Structural Analysis – SA
- Vocabulary – VO
- Sentence-Level Comprehension – SC
- Paragraph-Level Comprehension – PC
- Early Numeracy – EN

Students taking SEL answer specific numbers of items from each sub-domain; the SEL test blueprint specifies different numbers of those items, depending on a student's current grade and performance on previous administrations of SEL tests.

Star Early Literacy is an application of item response theory (IRT); specifically 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 early literacy skills in the case of Star Early Literacy.

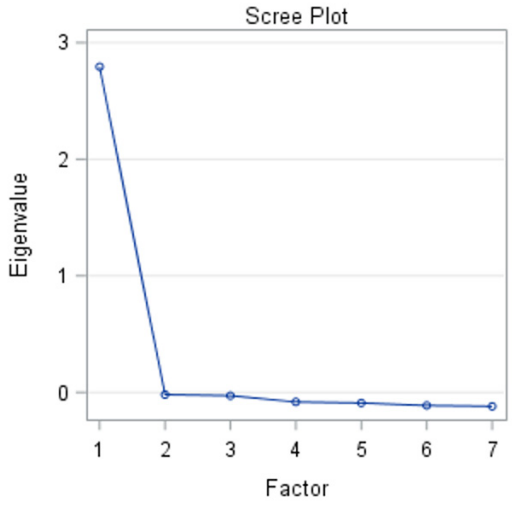
To evaluate whether Star Early Literacy measures a single construct, factor analyses were conducted. Factor analysis is a statistical technique used to determine the number of dimensions or constructs that a test measures. Both exploratory and confirmatory factor analyses were conducted across grades K to 3.

To begin, a large sample of student Star Early Literacy data was assembled. The overall sample consisted of 79,328 student records. That sample was divided into 2 sub-samples. The first sub-sample, consisting of 23,807 cases, was used for exploratory factor analysis; the second subsample, 55,521 cases, was reserved for confirmatory factor analyses that followed the initial exploratory analysis.

Within each sub-sample, each student's 27 Star Early literacy item responses were divided into subsets of items aligned to each of the 10 content sub-domains. Not all sub-domains are administered to all grades based on the expected developmental trajectory of students in the early grades. For example, the structural analysis and the paragraph-level comprehension domains are administered starting in grade 2 but advanced students in lower grades may be tested on these sub-domains, depending on their performance on prior administrations of the test. For each student, separate Rasch ability estimates (subtest scores) were calculated from each subdomain-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 1 to 6, following the Star Early Literacy test content blueprints, which specify different numbers of items per sub-domain, depending on the student's grade level. There are 5 Star Early Literacy blueprints for each grade level from Pre-K to grade 3.

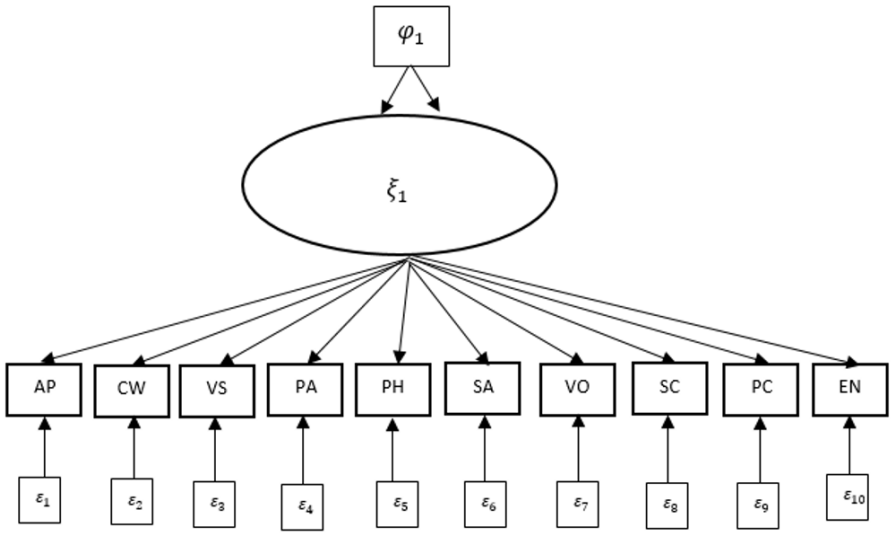
Intercorrelations of the subdomain-specific Rasch subtest scores based on the blueprint used were analyzed using exploratory factor analysis (EFA) to evaluate the number of dimensions/ factors underlying Star Early Literacy. Varimax rotation was used. For each blueprint used, the EFA analyses retained a single dominant underlying dimension based on the PROPORTION criterion (proportion of variance explained by the factor), as expected. An example of a scree plot from the kindergarten blueprint sample is shown in Figure 6. The other blueprint-specific EFA analyses had almost identical results to the Kindergarten blueprint sample.

Figure 6: Example Scree Plot from the Exploratory Factor Analysis in Star Early Literacy Using the Kindergarten Blueprint Sample



Subsequent to the EFA analyses, confirmatory factor analyses (CFA) were also conducted using the subtest scores from the CFA sub-sample. Five confirmatory analyses were conducted. The most complex CFA model tested a single underlying factor with all of the 10 sub-domains loading on it as shown in Figure 7. The names AP to EN in Figure 7 represent the ten Star Early Literacy sub-domains all loading on a single factor. This model applied to the blueprint samples for grades 2 and 3; the first grade blueprint model omitted the SA and the PC sub-domains; the kindergarten blueprint model also omitted the SA and PC sub-domains as well as the SC sub-domain; and the Pre-K blueprint model was similar to the Kindergarten model but further eliminated the PH sub-domain.

Figure 7: Confirmatory Factor Analyses (CFA) in Star Early Literacy



The results of the CFA analyses, grouped by the test blueprint administered, are summarized in Table 31. As that table indicates, the sample sizes were adequate for the CFA analyses. 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 all very close to 1 (or equal 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 Early Literacy.

Table 31: Summary of the Goodness-of-Fit of the CFA model for Star Early Literacy

Grade Blueprint	N	X^2	df	CFI	GFI	NFI	RMSEA	SRMR
Pre-K	2,249	43.5375	9	0.99	0.99	0.99	0.0413	0.019
K	15,424	229.057	14	0.99	1.00	0.99	0.0316	0.0139
1	11,904	1143.121	20	0.96	0.97	0.96	0.0687	0.0319
2	11,516	1679.2156	35	0.97	0.98	0.97	0.0639	0.0271
3	13,030	2618.7537	35	0.97	0.95	0.97	0.0753	0.0271

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).

Summary of Star Early Literacy Validity Data

In the aggregate, the data presented in the Validity section above are evidence of Star Early Literacy's concurrent, retrospective, predictive, and construct validity. The majority of the validity evidence presented in this chapter was specific to the versions of the assessment that preceded development of the current 27-item version. However, because of the similarity of the original and current versions, and the high degree of correlation between them reported here, there is ample reason to consider them equivalent, and therefore to accept evidence of the validity of those earlier versions as being applicable to the current version as well. Indeed, the observed correlation between the current and earlier version, corrected for attenuation due to measurement error, is nearly perfect. Along with the close correspondence of the current version's specific validity data to that of previous versions, this disattenuated correlation is evidence that the two versions are measuring a common underlying attribute, and doing so with equivalent degrees of measurement precision. We can confidently treat all of the evidence of the validity of the earlier Star Early Literacy as applying perforce to the current version of Star Early Literacy, and can accept all of the summary statements here as equally applicable to the current version.

As the data presented in this chapter attests, scores on Star Early Literacy increase systematically and substantially with age and school grade, reaching a plateau at grade 3, by which time the overwhelming majority of children have mastered the early literacy skills the test measures.

Scores on Star Early Literacy were also shown to be strongly related to teachers' ratings of children's skills, with easier skills mastered by children at

relatively low levels on the Star Early Literacy score scale, and more difficult skills mastered by children with scores at higher levels.

In concurrent test administrations, Star Early Literacy was found to correlate 0.78 with Star Reading in a sample of first- to third-grade students. Correlations with numerous other tests were summarized inTable 20. These summaries showed that Star Early Literacy correlated an average of 0.59 with a wide range of measures of early literacy, readiness, and reading administered in grades K through 3. In the Meta-Analysis section, Table 21 showed the average uncorrected correlation between Star Early Literacy and all of the other tests to be 0.60. (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 0.84.) Correlations with specific measures of reading ability were often higher than this average.

Research subsequent to publication shows relationships with other tests—including DIBELS, GRADE, the Michigan Literacy Progress Profile, Running Record, and TPRI—to be consistent with the Validity Study data: Star Early Literacy scores are moderately to highly correlated with scores on a wide range of measures of early literacy skills. Perhaps most importantly, research shows Star Early Literacy to be a valid predictor of children's later reading development, as measured by scores on reading tests administered two years later.

An exploratory factor-analytic study of the current 27-item version showed that even with the addition of a 5-item early numeracy section, the test continues to measure a single construct firmly related to early literacy. A separate confirmatory factor analysis of subscores based on items aligned to SEL's 10 sub-domains confirmed that the test as a whole is demonstrably unidimensional at each of the grades (K–3) at which SEL is typically used. This serves as evidence in support of the use of the Rasch IRT model as the foundation for both adaptive item selection and ability estimation (scoring) in Star Early Literacy.

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."

National norms for Star Early Literacy were first computed in 2015. These norms were used until 2017 when the Star Early Literacy norms were updated. This chapter describes the 2017 norming of Star Early Literacy.

Background

Previous Star Early Literacy Norms

The current version of Star Early Literacy, introduced in June 2011, is the first standards-based version of Star Early Literacy; it assesses a wide variety of skills and instructional standards, including sections on both early literacy and early numeracy skills. As part of its development, Star Early Literacy scale scores were equated to the scale used in earlier version of Star Early Literacy. The equating analyses demonstrated that, despite its distinctive content, the latent attribute underlying the current version is the same one underlying the previous version of Star Early Literacy. It measures the same broad construct, and reports student performance on the same score scale. Older norms included the 2014 and the 2015 Star Early Literacy norms. The 2015 norms will be replaced by the 2017 norms. Additionally for the 2017 Star Early Literacy norms, two reporting scales exist: One is the existing nonlinear transformation of the Star Early Literacy Rasch scores to scaled scores which will be referred to here as the Star Early Literacy scale scores; the other is a new linear transformation of the Star Early Literacy Rasch score to scaled scores which will be referred to as the Unified scale. Score summaries by grade will be presented in both scales.

The 2017 Star Early Literacy Norms

New US norms for Star Early Literacy assessments were introduced at the start of the 2017–18 school year. Separate early fall and late spring norms were developed for grades Kindergarten through 3. In previous Star Early

Literacy norming analyses, the reference populations for grades Kindergarten through 3 consisted only of students taking Star Early Literacy; students who only took Star Reading were excluded from the Star Early Literacy norms, and vice versa. Consequently, previous Star Early Literacy 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 Early Literacy and Reading 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.

Consistent with that concept, Renaissance Learning has developed a unified score scale that applies to both assessments. 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.

This chapter describes the development of the 2017 norms, which are based on user data collected over the course of one full school year: 2014–2015. Details of the procedures employed are given in this chapter.

Students participating in the norming study took assessments between August 15, 2014 and June 30, 2015. Students took both the Star Early Literacy and the Star Reading 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 and/or Star Early Literacy tests as they are administered in everyday use.

Sample Characteristics

During the norming period, a total of 941,126 US students in grades K–3 took current Star Early Literacy while 2,051,457 US students in the same grades took Star Reading tests, using Renaissance servers hosted by Renaissance Learning. The first step in sampling was to select a representative sample of students who had tested in the fall, in the spring, or in both the fall and spring of the 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 and Star Early Literacy customer base, steps were taken to ensure the resulting norms were nationally representative of grades K–3 US student populations with regard to certain important characteristics. A post-stratification procedure was used to adjust the samples 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 Early Literacy 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 1,893,599 students in grades K–3. There were 1,428,592 students in the fall norming sample and 998,490 students in the spring norming sample; 533,483 students were included in both norming samples. These students came from 13,388 schools across 50 states and the District of Columbia.

Table 32 provides a breakdown of the number of students participating per grade in the fall and in the spring samples.

Table 32: Numbers of Students per Grade in the Fall and in the Spring Samples

	Fall Sample	Spring Sample
Grade	N	N
K	212,035	196,720
1	340,079	237,360
2	456,566	264,790
3	419,912	299,620
Total	1,428,592	998,490

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–3 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 33 on page 85 shows national percentages of children in grades K–3 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 33: Sample Characteristics Along with National Population Estimates and Sample Estimates

		National Estimates	Fall Norming Sample	Spring Norming Sample
Region	Midwest	20.9%	20.2%	17.9%
	Northeast	19.2%	10.0%	11.8%
	Southeast	24.6%	37.0%	33.1%
	West	35.3%	32.8%	37.2%
School Enrollment	< 200	4.0%	3.5%	3.7%
	200–499	26.8%	41.4%	41.3%
	≥ 500	69.1%	55.1%	55.0%
District Socioeconomic Status	Low	19.5%	26.9%	28.6%
	Below Median	24.3%	30.0%	28.4%
	Above Median	25.2%	21.3%	21.2%
	High	31.1%	21.8%	21.8%
Location	Rural	14.1%	21.3%	20.3%
	Suburban	42.3%	37.1%	36.8%
	Town	11.7%	16.8%	17.8%
	Urban	31.9%	24.8%	25.1%

Table 34 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 and the sample summaries in Table 34 are based on only those students that had gender and ethnicity information available. In addition to the student demographics shown, an estimated 5.8% 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 Early Literacy 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 34: Student Demographics and School Information: National Estimates and Samples Percentages

			National Estimate	Fall Norming Sample	Spring Norming Sample
Gender	Public	Female	48.6%	49.7%	49.0%
		Male	51.4%	50.3%	51.0%
	Non-Public	Female	–	50.4%	50.3%
		Male	–	49.6%	49.7%
Race/Ethnicity	Public	American Indian	1.0%	1.5%	1.5%
		Asian	5.3%	4.7%	4.9%
		Black	15.5%	19.8%	19.5%
		Hispanic	25.4%	20.1%	23.1%
		White	49.6%	53.9%	51.0%
		Multiple Race[a]	3.2%	–	–
	Non-Public	American Indian	0.5%	3.0%	3.5%
		Asian	6.6%	2.8%	3.0%
		Black	9.1%	12.5%	15.1%
		Hispanic	10.7%	23.9%	28.5%
		White	69.2%	57.8%	49.9%
		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 users. Data were from a single school year from August 2014 to June 2015. Students' Rasch scores on their first Star Reading or Early Literacy test taken between the first and the second month 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 between the 8th and the 9th month 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 Rasch ability estimate. Norms were developed based on the weighted Rasch ability estimates and then transformed to unified as well as Star Early Literacy scaled scores.[2] Table 35 provides descriptive statistics for each grade with respect to the normative sample performance, in the Unified scaled score units. Table 5 provides descriptive statistics for each grade with respect to the normative sample performance, in the Star Early Literacy scaled score units.

Table 35: Descriptive Statistics for Weighted Scaled Scores by Grade for the Norming Samples in the Unified Scale

	Fall Unified Scaled Scores				Spring Unified Scaled Scores			
Grade	N	Mean	Standard Deviation	Median	N	Mean	Standard Deviation	Median
K	212,035	702	62	703	196,720	796	65	793
1	340,079	776	72	767	237,360	857	69	856
2	456,566	887	70	888	264,790	939	66	942
3	419,912	952	67	956	299,620	987	64	990

Table 36: Descriptive Statistics for Weighted Scaled Scores by Grade for the Norming Samples in the Star Early Literacy Scale

	Fall Star Early Literacy Scaled Scores				Spring Star Early Literacy Scaled Scores			
Grade	N	Mean	Standard Deviation	Median	N	Mean	Standard Deviation	Median
K	212,035	521	104	522	196,720	692	107	687
1	340,079	657	123	641	237,360	782	83	781
2	456,566	805	79	806	264,790	850	45	852
3	419,912	846	50	849	299,620	866	32	867

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.

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 Early Literacy 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 Early Literacy 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

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

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 Early Literacy SGP growth norms are based on a sample of more than 933,000 student records across grades K–3, with grade-specific samples ranging from about 17,000 to 494,000.

Score Definitions

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

For its internal computations, Star Early Literacy uses procedures associated with the Rasch 1-parameter logistic response model. A proprietary Bayesian-modal item response theory estimation method is used for scoring until the student has answered at least one item correctly and at least one item incorrectly. Once the student has met this 1-correct/1-incorrect criterion, Star Early Literacy software uses a proprietary Maximum-Likelihood IRT estimation procedure to avoid any potential bias in the Scaled Scores. All Star Early Literacy item difficulty values are Rasch model parameters.

Adaptive item selection is predicated on matching Rasch item difficulty and ability parameters, and students' abilities are expressed on a Rasch scale. For score reporting purposes, however, transformed scores are used. Three kinds of transformations of the Rasch ability scale are used: Scaled Scores, criterion-referenced scores, and norm-referenced scores.

Types of Test Scores

Star Early Literacy software provides three broad 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. On the basis of Scaled Scores, students taking Star Early Literacy are categorized into one of three literacy classifications (see the Literacy Classification section below). In addition, Star Early Literacy uses two types of proficiency scores: Sub-domain Scores and Skill Set Scores, and Estimated Oral Reading Fluency scores (Est. ORF).
- *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 Early Literacy 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 scores* are the fundamental scores used to summarize students' performance on Star Early Literacy tests. Upon completion of the test, the testing software calculates a single-valued Star Early Literacy Scaled Score as well as a unified scale score. The Scaled Score is a non-linear, monotonic transformation of the Rasch ability estimate resulting from the adaptive test. The unified scale score is a linear transformation of the Rasch estimate, as described below.

Star Early Literacy 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 Early Literacy tests to scores on a common scale. Star Early Literacy 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 Scaled Scores.

Two different scaled scores are now available: the original scale scores, which are referred to as Star Early Literacy scale scores; and a new score, expressed on the "Unified" score scale, which was introduced with the 2017–2018 school year.

Star Early Literacy Scale Scores

The Star Early Literacy scale scores are the same scores that have been reported continuously since Star Early Literacy Version 1 was introduced, in 2001.They range from 300 to 900. The extreme scores 300 and 900, although possible, occur rarely.

Unified Scale Scores

Many users of Star Early Literacy use it to assess their students until they are ready to take Star Reading. 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 1999 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 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 on that scale; Star Reading Unified scale scores range from 600 to 1400. The overlap between those ranges is the range of development in which students can be assessed with either or both of the two Star tests. 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 Star Early Literacy scale scores.

Sub-domain and Skill Set Scores

Star Early Literacy uses proficiency scores to express a student's expected performance in the ten sub-domains and 41 subordinate skill sets that make up the Star Early Literacy item bank. These proficiency scores are referred to in Star Early Literacy score reports as Sub-domain Scores and Skill Set Scores. Each Sub-domain Score is a statistical estimate of the percent of items the student would be expected to answer correctly if all of the Star Early Literacy items in the sub-domain were administered. Therefore, Sub-domain Scores range from 0 to 100 percent.

Similarly, a Skill Set Score estimates the percent of all the Star Early Literacy items in a specific skill set that the student would be expected to answer correctly. Sub-domain and Skill Set Scores are calculated by applying the Rasch model. The student's measured Rasch ability, along with the known Rasch difficulty parameters of the items within the appropriate sub-domain or skill, are used to calculate the expected performance on every item. The average expected performance on the items that measure a given sub-domain or skill set is used to express each Sub-domain or Skill Set Score.

Literacy Classification

Star Early Literacy score reports include a classification of the student into one of three literacy classifications or reading development stages, based on the Scaled Score. Students with Star Early Literacy Scaled Scores below 675 are classified as "Emergent Readers," those with scores from 675 through 774

are classified as "Transitional Readers," and those scoring 775 and above are classified as "Probable Readers."

The cut points for these three categories are competency-based. To be classified as a Transitional Reader, a student needs to have mastered specific skills that are represented in the Star Early Literacy item bank. Similarly, to be classified as a Probable Reader, mastery of higher-level skills must be apparent. Table 43 and Table 44 in the Conversion Tables section contains the literacy classifications in both the Star Early Literacy and the Unified scales.

Estimated Oral Reading Fluency (Est. ORF)

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

Est. ORF is reported as the estimated number of words a student can read correctly within a one-minute time span on grade-level-appropriate text. Grade-level text is 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, the score interpretation for a second-grade student with an Est. ORF score of 60 would be that the student is expected to read 60 words correctly within one minute on a passage with a readability level between 2.0 and 2.5. Therefore, when this estimate is compared to observed scores, there might be noticeable differences, as the Est. ORF provides an estimate across a range of readability but an individual oral reading fluency passage would have a fixed level of difficulty. See Table 45 in the Conversion Tables section for the current Est. ORF table for Star Early Literacy.

Percentile Ranks (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 Early Literacy or 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.

Beginning with the 2017–2018 school year Star Early Literacy and Star Reading share a single set of test score norms, based on the joint distribution of their unified scale scores. Table 39 and Table 40 in the Conversion Tables chapter contain an abridged versions of both the Star Early Literacy and the Unified Scaled Scores to Percentile Rank conversion table that the Star software uses to look up percentile ranks for each test. The unabridged table includes data for all of the monthly grade placement values from 0.0–3.9 (Kindergarten through grade 3.)

Normal Curve Equivalent (NCE) Scores

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 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 41 in the Conversion Tables chapter lists the NCEs corresponding to integer PR values and facilitates the conversion of PRs to NCEs. Table 42 provides the conversions from NCE to PR. The NCE values are given as a range of scores that convert to the corresponding PR value.

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 2.4, this means that the student scored as well on Star Early Literacy as did the typical student in the fourth month of grade 2. It does not necessarily mean that the student can read independently at a second-grade level, only that he or she obtained a Scaled Score as high as the average second-grade, fourth-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 Early Literacy Grade Equivalents range from 0.0–3.9+. 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 37 indicates how the GE scale corresponds to the various calendar months. For example, if a student obtained a GE of 3.6 on a Star Early Literacy assessment, this would suggest that the student was performing similarly to the average student in the third grade at the sixth month (March) of the academic year. Because Star Early Literacy 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 38 contains the Scaled Score to GE conversions.

Table 1: 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 Star Early Literacy with Conventional Tests

Because the Star Early Literacy test adapts to the reading level of the student being tested, Star Early Literacy 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 first-grade student to obtain a GE score of 3.9 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 third-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 first-grade test.

Star Early Literacy 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 3.6 indicates that the student's score can be appropriately compared to that of a typical third-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 third-grade reading material).

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 testing windows, and
- took the most recent test and the first prior test on the same dates.

SGPs provide a measure of how a student changed from one Star testing window[1] to the next relative to other students with similar starting Star Early

1. We collect data for our growth norms during three different time periods: fall, winter, and spring.

Literacy 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 Early Literacy score history. All students, no matter their starting Star score, have an equal chance to demonstrate growth at any of the 99 percentiles.

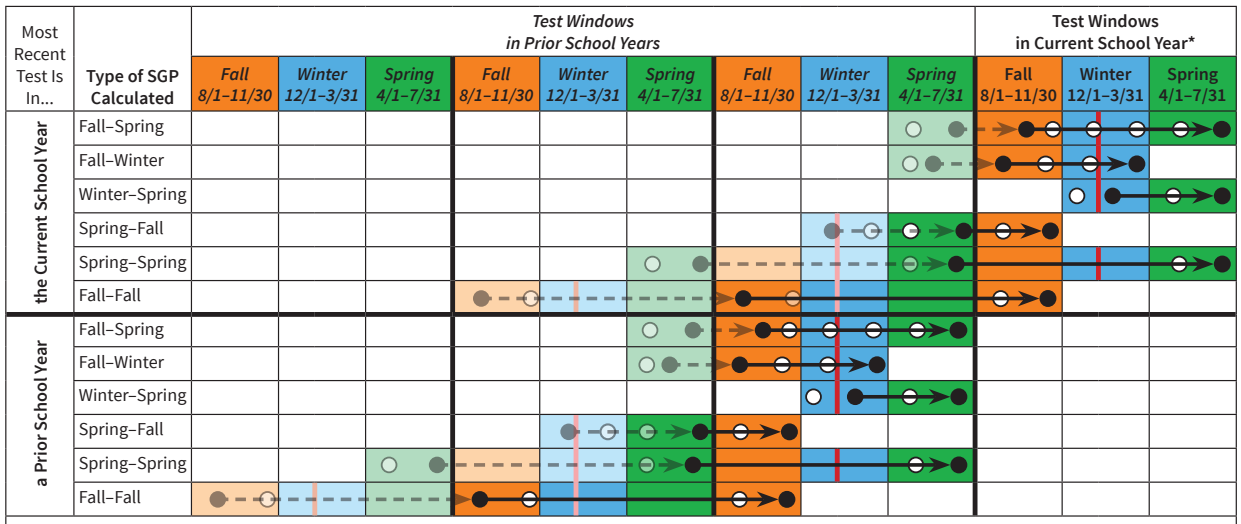
SGPs are often used to indicate whether a student's growth is more or less than can be expected. For example, without an SGP, a teacher would not know if a Scaled Score increase of 100 points represents good, not-so-good, or average growth. This is because students of differing achievement levels in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

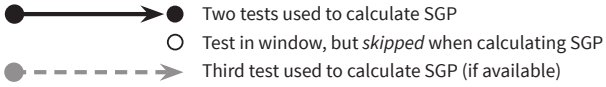
SGP is calculated for students who have taken at least two tests (a current test and a prior test) within at least two different testing windows (Fall, Winter, or Spring).

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 and Lexile® Ranges

In cooperation with MetaMetrics®, in the winter of 2018–2019, users of Star Early Literacy were given the option of including Lexile® Measures on certain Star Early Literacy score reports. Reported Lexile® Measures range from BR1300L to 2220L. (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® Ranges are calculated from Lexile® Measures. Lexile® Ranges are calculated by subtracting 100L from and adding 50L to a student's Lexile® Measure. For example, a student with a Lexile® Measure of 700L would have a Lexile® Range of 600L–750L. In Star Early Literacy, Lexile® Measures below BR400L are shown on reports for progress monitoring purposes only. A score below BR400L is not meant to be used to match readers with text; in these cases, a Lexile® Range is not reported.

Conversion Tables

Table 38: Star Early Literacy (SEL) Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score		Star Early Literacy Scaled Score	
	Low	High	Low	High
0	200	698	300	513
0.1	699	708	514	531
0.2	709	718	532	549
0.3	719	728	550	568
0.4	729	738	569	586
0.5	739	748	587	605
0.6	749	757	606	622
0.7	758	767	623	641
0.8	768	776	642	657
0.9	777	785	658	673
1	786	794	674	689
1.1	795	802	690	703
1.2	803	811	704	717
1.3	812	819	718	730
1.4	820	827	731	742
1.5	828	835	743	753
1.6	836	843	754	764
1.7	844	851	765	774
1.8	852	859	775	784
1.9	860	866	785	792
2	867	874	793	800
2.1	875	881	801	807
2.2	882	888	808	814
2.3	889	895	815	820
2.4	896	902	821	826
2.5	903	908	827	831

Table 38: Star Early Literacy (SEL) Scaled Score to Grade Equivalent Conversions

Grade Equivalent	Unified Scaled Score		Star Early Literacy Scaled Score	
	Low	High	Low	High
2.6	909	915	832	836
2.7	916	921	837	840
2.8	922	927	841	844
2.9	928	934	845	848
3	935	940	849	852
3.1	941	946	853	855
3.2	947	951	856	857
3.3	952	957	858	860
3.4	958	963	861	863
3.5	964	968	864	865
3.6	969	973	866	867
3.7	974	978	868	869
3.8	979	983	870	870
3.9	984	988	871	871
4	989	1100	872	900

Table 39: Star Early Literacy Scaled Score to Percentile Rank Conversions[a]

	Grade (First Month)					Grade (First Month)					Grade (First Month)					Grade (First Month)			
PR	K	1	2	3	PR	K	1	2	3	PR	K	1	2	3	PR	K	1	2	3
1	300	300	300	300	26	460	574	736	809	51	523	644	807	850	76	593	734	850	872
2	358	412	554	617	27	463	576	741	812	52	527	646	809	—	77	597	738	852	873
3	367	429	576	649	28	466	580	744	814	53	528	649	812	852	78	601	743	853	—
4	374	444	593	671	29	469	582	747	817	54	530	653	813	853	79	604	747	854	874
5	379	457	608	689	30	472	586	751	819	55	534	656	815	—	80	608	753	856	875
6	386	467	620	704	31	475	587	755	821	56	536	660	817	855	81	612	757	857	—
7	389	476	630	719	32	478	589	758	823	57	539	664	819	—	82	616	762	858	876
8	395	484	641	729	33	481	593	761	824	58	541	667	820	856	83	621	768	860	877
9	399	492	649	739	34	484	597	764	826	59	543	671	822	857	84	625	773	861	—
10	405	501	657	748	35	486	599	768	828	60	547	674	824	858	85	631	776	863	878
11	408	508	665	755	36	489	601	771	830	61	548	678	826	859	86	636	781	864	879
12	412	513	671	760	37	492	604	774	831	62	552	681	828	860	87	644	786	865	880

Table 39: Star Early Literacy Scaled Score to Percentile Rank Conversions[a]

	Grade (First Month)					Grade (First Month)					Grade (First Month)					Grade (First Month)			
PR	K	1	2	3	PR	K	1	2	3	PR	K	1	2	3	PR	K	1	2	3
13	417	518	678	764	38	494	608	777	833	63	554	685	830	861	88	649	790	867	881
14	420	523	684	769	39	497	610	779	834	64	558	688	831	862	89	656	795	868	—
15	423	528	688	773	40	499	612	782	836	65	561	692	833	863	90	665	800	870	882
16	428	534	694	779	41	502	616	785	837	66	563	695	835	—	91	673	805	871	883
17	432	539	699	783	42	504	617	787	838	67	567	699	836	864	92	681	811	873	—
18	435	543	703	787	43	506	621	790	840	68	569	702	838	865	93	690	818	874	884
19	437	547	709	789	44	509	623	792	841	69	572	705	840	866	94	697	824	876	885
20	440	552	712	792	45	511	625	793	843	70	576	709	841	867	95	707	830	878	886
21	444	556	717	796	46	513	629	796	—	71	578	712	843	868	96	720	838	880	887
22	448	559	723	798	47	514	631	798	845	72	582	715	844	869	97	737	845	881	888
23	450	563	726	801	48	516	634	800	846	73	586	720	846	—	98	760	854	884	890
24	452	567	729	804	49	520	636	803	847	74	587	725	847	870	99	789	866	887	892
25	455	570	733	807	50	521	640	805	848	75	591	729	848	871					

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

Table 40: SR and SEL Unified Scaled Score to Percentile Rank Conversions[a]

	Grade (First Month)					Grade (First Month)					Grade (First Month)					Grade (First Month)			
PR	K	1	2	3	PR	K	1	2	3	PR	K	1	2	3	PR	K	1	2	3
1	200	200	200	200	26	665	732	837	911	51	704	769	889	958	76	743	822	939	1002
2	568	629	743	797	27	667	733	840	913	52	706	770	891	959	77	744	825	942	1004
3	581	643	754	812	28	669	735	842	915	53	707	772	893	961	78	746	828	944	1006
4	589	654	763	823	29	671	736	844	918	54	708	774	894	963	79	748	831	946	1008
5	595	663	770	832	30	673	738	846	920	55	710	776	896	964	80	750	834	949	1010
6	602	670	776	840	31	675	739	849	922	56	711	778	898	966	81	752	838	951	1012
7	607	676	781	847	32	677	740	851	924	57	713	780	900	967	82	755	842	954	1014
8	613	681	786	854	33	679	742	853	926	58	714	782	902	969	83	757	846	956	1016
9	617	686	790	860	34	681	744	856	928	59	715	784	904	971	84	759	849	959	1018
10	622	691	794	865	35	682	745	858	930	60	717	786	906	972	85	762	853	962	1021
11	625	695	798	869	36	684	746	860	932	61	718	788	908	974	86	765	857	965	1024
12	629	698	801	873	37	686	748	862	934	62	720	790	910	975	87	769	861	968	1027
13	633	701	805	876	38	687	750	864	935	63	721	792	912	977	88	772	865	971	1029
14	635	704	808	879	39	689	751	866	937	64	723	794	914	979	89	776	869	974	1032
15	638	707	810	882	40	690	752	868	939	65	725	796	916	981	90	781	874	978	1035
16	642	710	813	886	41	692	754	870	941	66	726	798	918	982	91	785	879	982	1038
17	645	713	816	889	42	693	755	872	942	67	728	800	920	984	92	790	885	986	1041
18	647	715	818	892	43	694	757	874	944	68	729	802	923	986	93	795	892	990	1044

Table 40: SR and SEL Unified Scaled Score to Percentile Rank Conversions[a]

	Grade (First Month)					Grade (First Month)					Grade (First Month)					Grade (First Month)			
PR	K	1	2	3	PR	K	1	2	3	PR	K	1	2	3	PR	K	1	2	3
19	649	717	821	894	44	695	758	876	946	69	731	804	925	988	94	799	899	995	1048
20	651	720	823	896	45	697	759	877	948	70	733	806	927	990	95	805	907	1001	1053
21	654	722	826	899	46	698	761	879	949	71	734	808	929	992	96	813	917	1008	1058
22	657	724	829	901	47	699	762	881	951	72	736	810	931	994	97	824	929	1015	1066
23	659	726	831	904	48	700	764	883	953	73	738	813	933	996	98	840	943	1025	1075
24	660	728	833	906	49	702	765	885	954	74	739	816	935	998	99	865	969	1041	1091
25	662	730	835	909	50	703	767	887	956	75	741	819	937	1000					

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

Table 41: Percentile Rank to Normal Curve Equivalent Conversions

PR	NCE	PR	NCE	PR	NCE	PR	NCE
1	1	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
5	15.4	30	39	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	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	62	56.4	87	73.7
13	26.3	38	43.6	63	57	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
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	95	84.6
21	33	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
25	35.8	50	50	75	64.2		

Table 42: 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	4	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	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	80
16.3	18	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	8	40.5	40.9	33	54	54.4	58	69.7	70.4	83
21.1	22.3	9	41	41.5	34	54.5	55	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	45.4	41	58.4	58.9	66	77.6	78.8	91
29.5	30.2	17	45.5	45.9	42	59	59.5	67	78.9	80.2	92
30.3	31	18	46	46.5	43	59.6	60.1	68	80.3	81.7	93
31.1	31.8	19	46.6	47	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	96
33.4	34	22	48.2	48.6	47	62	62.5	72	88.1	91	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	99
35.5	36	25	49.8	50.2	50	63.9	64.5	75			

Table 43: Relating Star Early Literacy Scale Scores to Star Reading Enterprise Scores

Star Early Literacy			Star Reading				
Star Early Literacy Scaled Score Range		Literacy Classification	SR 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
383	513		0	54	0.0	0.0–1.0	
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 SS = 675	81	83	1.0	1.0–2.0	
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 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	
841	844		286	298	2.8	2.5–3.5	
845	848		299	314	2.9	2.5–3.5	
849	852		315	326	3.0	2.6–3.6	

Table 43: Relating Star Early Literacy Scale Scores to Star Reading Enterprise Scores

Star Early Literacy			Star Reading				
Star Early Literacy Scaled Score Range		Literacy Classification	SR Enterprise Scaled Score Range		GE	ZPD Range	Recommended Assessment(s)
Low	High		Low	High			
853	855	Probable Reader (continued)	327	340	3.1	2.6–3.7	Star Reading (continued)
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 44: Relating Star Early Literacy Unified Scale Scores to Star Reading GE Scores and ZPD Ranges

Star Early Literacy			Star Reading		
SEL Literacy Classification	Unified Scaled Score Range		GE	ZPD Range	Recommended Assessment(s)
	Low	High			
Emergent Reader	NA	NA	NA	NA	Star Early Literacy
	600	698	0.0	0.0–1.0	
	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 SS = 786	786	794	1.0	1.0–2.0	
	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	

Table 44: Relating Star Early Literacy Unified Scale Scores to Star Reading GE Scores and ZPD Ranges

Star Early Literacy			Star Reading		
SEL Literacy Classification	Unified Scaled Score Range		GE	ZPD Range	Recommended Assessment(s)
	Low	High			
Transitional Reader (continued)	828	835	1.5	1.5–2.5	Star Early Literacy and Star Reading
	836	843	1.6	1.6–2.6	
	844	851	1.7	1.7–2.7	
Probable Reader 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	
	922	927	2.8	2.5–3.5	
	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	

Table 45: Estimated Oral Reading Fluency (Est. ORF) Given in Words Correct per Minute (WCPM) by Grade for Selected Star Early Literacy Scale Score Units (SEL SS)

Star Early Literacy SS	Unified SS	Grade		
		1	2	3
300–460	200–665	0	0	0
480	678	0	1	0
500	690	0	3	0
520	702	1	6	3
540	713	5	7	6
560	724	8	9	8
580	735	11	13	10
600	745	13	15	12
620	756	16	18	16
640	767	18	20	18
660	777	21	22	21
680	789	24	25	25
700	800	26	27	28
720	813	29	30	32
740	826	34	34	37
760	840	42	40	44
780	856	51	49	50
800	873	62	58	56
820	894	74	71	65
840	920	88	87	78
860	955	140	112	103
880	1013	142	175	150
900	1100	142	175	170

Appendix: Detailed Validity Data

Relationship of Star Early Literacy Scores to Other Tests

This appendix contains selected, detailed data from two validity studies conducted prior to the 2001 publication of Version 1 of Star Early Literacy. These will be referred to as the calibration study and the validation study.

The calibration study, conducted in the fall of 2000, involved administration of grade-leveled fixed-form tests to a large sample of students, for the purpose of collecting item response data on the then uncalibrated items developed for Star Early Literacy. A total of 32,493 students took one or more of 246 test forms consisting of 40 test items. Following the completion of Rasch model calibration of the items, each student's Rasch ability score was calculated. Correlations between those scores and available measures on the same students were calculated, for use in early evaluation of the correlational validity of Star Early Literacy.

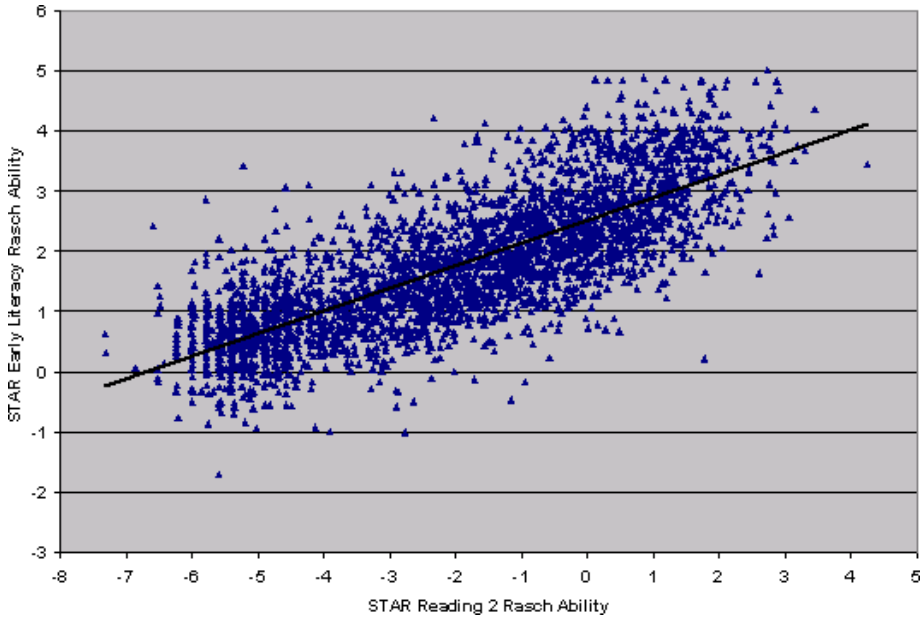
The validation study, conducted in early 2001, involved the first adaptive administration of Star Early Literacy, for the purpose of evaluating the validity and other characteristics of the test. As part of that evaluation, correlations of scores from the validation study and a variety of standardized reading tests were calculated.

Correlational results from both the 2000 calibration study and the 2001 validation study are presented here, in somewhat more detail than what is contained in the Validity chapter of the main body of this technical manual. The intent or recording these data here is to help maintain the historical record of early correlational research on Star Early Literacy.

Calibration Study Results

During the Calibration Study, over 3,000 students in grades 1 through 3 took Star Reading version 2 in addition to the Star Early Literacy tests. Figure 8 shows a plot of Star Early Literacy Rasch ability scores against Star Reading 2 Rasch ability scores.

Figure 8: Scatterplot of Star Early Literacy and Star Reading Rasch Ability Scores (N = 3,043 Students; Correlation = 0.78)



As the shape of the scatterplot suggests, the degree of correlation was substantial: overall, the correlation between Star Early Literacy scores and Star Reading scores was 0.78. This suggests that there is a strong relationship between the literacy skills measured by Star Early Literacy and reading proficiency as measured by Star Reading. Because the contents and formats of these two tests are quite dissimilar, the high degree of correlation between them supports the position that Star Early Literacy measures skills that are highly related to the development of reading ability in the early grades.

Validation Study and Later Study Data

In the Validation Study, participating teachers were asked to provide students' scores from a variety of other tests. In addition to reading test scores, scores on a number of other tests were obtained for many of the students participating in the Validation Study. These tests included other measures of early literacy as well as tests of readiness, social skills, and other attributes.

Usable scores were received for over 2,400 students on 20 different test series administered in the fall or spring of the 2000 school year or the spring of 2001. Most of the reported scores were either NCE scores or Scaled Scores. In a few cases, letter grades were reported; these were recoded into numbers in order to perform correlation analyses. From the usable data, 61 correlations with Star Early Literacy were computed. The number of

correlations ranged from 10 at the kindergarten level to 22 at grade 3. No external test scores were reported for pre-kindergarten students, with the exception of a subsequent predictive validation study.

As part of the ongoing efforts to provide evidence for the validity of Star Early Literacy scores, further research studies have been carried out. Additional concurrent validity studies have been undertaken, and the results were added to the overall results. Concurrent validity was operationally defined as the extent to which Star Early Literacy scores correlated with scores on external measures, and both tests were given within the same two-month period.

In addition, predictive validity studies were undertaken to provide some measure of the utility of using Star Early Literacy for predicting later outcomes. Predictive validity was defined as the extent to which scores on the Star tests predict 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. It provided an estimate of the linear relationship between Star scores and scores on measures 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 46 and Table 47 present the correlation coefficients between the scores on Star Early Literacy and each of the other test instruments (external measures) for which data were received. Table 46 displays "concurrent validity" data, that is, correlations observed when two test scores and other tests were administered at close to the same time. Table 47 provides validity estimates with external tests given prior to Star Early Literacy administration in spring 2001. Table 48 provides the predictive validity estimates with criterion tests given well after Star Early Literacy.

Table 46, Table 47, and Table 48 include the names of the external tests, the form or edition where known, the score type, the sample sizes (n), and correlations (r) computed at each of the four grades where data were reported. Averages of the correlations were calculated overall and by grade.

The averages of the concurrent validity correlations within grade were 0.64, 0.68, 0.52, and 0.57 for grades K–3 respectively. The overall concurrent correlation was 0.59.

The averages of the other correlations within grade were 0.49, 0.63, 0.57, and 0.59 for grades K–3 respectively. The average correlation was 0.58.

The average predictive validity coefficients for pre-K–3 were, respectively, 0.57, 0.52, 0.62, 0.67, and 0.77. The overall average predictive validity coefficient across the grades was 0.58.

Table 46: Concurrent Validity: Star Early Literacy Correlations with Tests Administered in Spring 2001, Grades K–3[a]

Test Form	Date	Score	K		1		2		3	
			n[b]	r	n	r	n	r	n	r
Brigance K & 1 Screen for Kindergarten and First Grade Children										
Revised	Spring 01	Scaled	21	0.64*	–	–	–	–	–	–
Revised	Spring 01	Scaled	19	0.61*	–	–	–	–	–	–
Canadian Achievement Test										
2nd Ed	Spring 01	Scaled	–	–	–	–	–	–	19	0.88*
Child Observation Record (COR)										
PC	Spring 01	NCE	–	–	–	–	83	0.67*	–	–
PC	Spring 01	Scaled	–	–	–	–	18	0.45	–	–
Developing Skills Checklist (DSC)										
	Spring 01	NCE	72	0.70*	–	–	–	–	–	–
Developmental Indicators for the Assessment of Learning (DIAL)										
3rd Ed	Spring 01	Scaled	–	–	–	–	50	0.42*	–	–
Florida Comprehensive Assessment Test (FCAT)										
	Spring 01	NCE	–	–	–	–	–	–	23	0.28
Gates-MacGinitie Reading Test (GMRT)										
Fourth S	Spring 01	NCE	–	–	–	–	12	0.76*	18	0.74*
2nd Can, A	Spring 01	Scaled	–	–	23	0.60*	–	–	–	–
2nd Can, B4	Spring 01	Scaled	–	–	–	–	24	0.34	–	–
2nd Can, C4	Spring 01	Scaled	–	–	–	–	–	–	11	0.54
Iowa Test of Basic Skills (ITBS)										
Form M	Spring 01	NCE	–	–	–	–	66	0.46*	80	0.54*
Unknown	Spring 01	NCE	–	–	63	0.72*	–	–	–	–
Form M	Spring 01	Scaled	–	–	–	–	13	0.53	–	–
Metropolitan Early Childhood Assessment Program (MKIDS)										
	Spring 01	NCE	14	0.88*	–	–	–	–	–	–
Stanford Achievement Test										
9th Ed	Spring 01	NCE	–	–	46	0.52*	21	0.50*	62	0.60*
9th Ed	Spring 01	Scaled	–	–	38	0.55*	38	0.79*	28	0.65*
STAR Reading										
Version 2	Spring 01	NCE	–	–	85	0.68*	69	0.39*	–	–
Version 2	Spring 01	Scaled	–	–	–	–	98	0.64*	117	0.57*

Table 46: Concurrent Validity: Star Early Literacy Correlations with Tests Administered in Spring 2001, Grades K–3[a]

Test Form	Date	Score	K		1		2		3	
			n[b]	r	n	r	n	r	n	r
TerraNova										
	Spring 01	NCE	–	–	6	0.95*	–	–	–	–
	Spring 01	Scaled	–	–	–	–	–	–	26	0.34
Test of Phonological Awareness (TOPA)										
	Spring 01	Scaled	11	0.68*	–	–	–	–	–	–
Texas Primary Reading Inventory (TPRI)										
	Spring 01	Letter	61	0.33*	–	–	–	–	–	–

Summary					
Grade(s)	**All**	**K**	**1**	**2**	**3**
Number of students	1,376	198	281	513	384
Number of coefficients	34	6	7	12	9
Average validity	–	0.64	0.68	0.52	0.57
Overall average	0.59				

a. No external test scores were reported for pre-kindergarten students.
b. Sample sizes are in the columns labeled "n" and correlation coefficients are in the columns labeled "r."* Denotes that a correlation is statistically significant at the 0.05 level.

Table 47: Other External Validity Data: Star Early Literacy Correlations with Tests Administered Prior to Spring 2001, Grades K–3[a]

Test Form	Date	Score	K		1		2		3	
			n[b]	r	n	r	n	r	n	r
Alabama Early Learning Inventory										
	Fall 00	Letter	32	0.42*	–	–	–	–	–	–
Gates-MacGinitie Reading Test (GMRT)										
Fourth	Spring 00	NCE	–	–	55	0.62*	–	–	–	–
Fourth S	Fall 00	NCE	–	–	59	0.60*	–	–	–	–
Indiana Statewide Testing for Educational Progress (ISTEP)										
	Fall 00	NCE	–	–	–	–	–	–	52	0.79*
Iowa Test of Basic Skills (ITBS)										
Form K	Spring 00	NCE	–	–	–	–	29	0.67*	39	0.57*
Form K	Fall 00	Scaled	–	–	30	0.56*	–	–	43	0.61*
Form M	Fall 00	Scaled	–	–	–	–	–	–	28	0.49*
Kaufman Survey of Early Academic and Language Skills (K-SEALS)										
1993	Fall 00	NCE	24	0.22	–	–	–	–	–	–

Table 47: Other External Validity Data: Star Early Literacy Correlations with Tests Administered Prior to Spring 2001, Grades K–3[a]

Test Form	Date	Score	K		1		2		3	
			n[b]	r	n	r	n	r	n	r
Metropolitan Readiness Test (MRT)										
6 Ed, Lev2	Spring 00	NCE	–	–	12	0.61	–	–	–	–
NWEA Levels Test										
	Fall 00	Scaled	–	–	–	–	–	–	48	0.51*
Stanford Achievement Test										
9th Ed	Spring 00	NCE	–	–	–	–	24	0.71*	80	0.49*
9th Ed	Spring 00	Scaled	–	–	–	–	61	0.47*	48	0.55*
9th Ed	Fall 00	NCE	25	0.85*	–	–	53	0.52*	63	0.73*
Stanford Test of Academic Skills										
	Fall 00	Scaled	–	–	–	–	–	–	27	0.71*
Star Reading										
Version 2	Winter 01	Scaled	–	–	20	0.75*	21	0.31	–	–
Version 2	Fall 00	Scaled	–	–	–	–	–	–	13	0.71*
TerraNova										
	Spring 00	Scaled	–	–	–	–	69	0.64*	68	0.62*
	Spring 00	Scaled	–	–	–	–	–	–	17	0.46
	Fall 00	Scaled	–	–	–	–	38	0.70*	31	0.44*
Texas Primary Reading Inventory (TPRI)										
	Fall 00	Letter	13	0.46	–	–	–	–	–	–

Summary					
Grade(s)	**All**	**K**	**1**	**2**	**3**
Number of students	1,122	94	176	295	557
Number of coefficients	29	4	5	7	13
Average validity	–	0.49	0.63	0.57	0.59
Overall average	0.58				

a. No external test scores were reported for pre-kindergarten students.
b. Sample sizes are in the columns labeled "n" and correlation coefficients are in the columns labeled "r."
* Denotes that a correlation is statistically significant at the 0.05 level.

Table 48: Predictive Validity Data: Star Early Literacy Predicting Later Performance for Grades Pre-K–3

Predictor Date	Criterion Date[a]	Pre-K[b]		K		1		2		3	
		n	r	n	r	n	r	n	r	n	r
Star Early Literacy											
Fall 05	Spr 06	142	0.47*	7,091	0.53*	7,394	0.61*	1,361	0.69*	201	0.76*
Fall 06	Spr 07	371	0.61*	10,231	0.51*	9,174	0.62*	1,704	0.73*	357	0.77*
Fall 05	Fall 06P	–	–	1,945	0.47*	685	0.64*	30	0.90*	–	–
Fall 05	Spr 07P	–	–	1,945	0.42*	685	0.62*	30	0.72*	–	–
Spr 06	Fall 06P	22	0.67*	1,945	0.58*	685	0.77*	30	0.85*	–	–
Spr 06	Spr 07P	22	0.50*	1,945	0.59*	685	0.71*	30	0.71*	–	–
Star Reading											
Fall 03	Fall 05P	–	–	671	0.49*	698	0.58*	194	0.65*	–	–
Win 04	Fall 05P	–	–	671	0.54*	698	0.62*	194	0.61*	–	–
Spr 04	Fall 05P	–	–	671	0.73*	698	0.67*	194	0.65*	–	–
Fall 03	Win 06P	–	–	552	0.43*	653	0.56*	469	0.64*	–	–
Win 04	Win 06P	–	–	858	0.55*	772	0.61*	227	0.57*	–	–
Spr 04	Win 06P	–	–	639	0.51*	551	0.66*	254	0.59*	–	–
Fall 03	Spr 06P	–	–	282	0.47*	376	0.61*	291	0.62*	–	–
Win 04	Spr 06P	–	–	497	0.56*	428	0.59*	167	0.59*	–	–
Spr 04	Spr 06P	–	–	480	0.55*	343	0.58*	195	0.57*	–	–

Summary						
Grade(s)	**All**	**Pre-K**	**K**	**1**	**2**	**3**
Number of students	61,443	557	30,423	24,525	5,370	558
Number of coefficients	51	4	15	15	15	2
Average validity	–	0.57	0.52	0.62	0.67	0.77
Overall average	0.58					

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

References

Adams, M. J. (1990). *Beginning to read.* London: MIT Press.

Anderson, R. C. (1996). *Research foundations to support wide reading.* Publication of the Center for the Study of Reading, Technical Report no. 631. Champaign, IL: University of Illinois at Urbana–Champaign.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report on the commission of reading.* Washington, DC: The National Institute of Education.

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly, 23,* 285–303.

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

Betts, J. E. & McBride, J. R. (2006). *Validity evidence for a computer-adaptive measure of early literacy skills for progress monitoring and prediction of later reading achievement.* Internal technical report, Renaissance Learning, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16, 297–334.

Deno, S. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37,* 184–192.

Dukette, D. & Cornish, D. (2009). *The essential 20: Twenty components of an excellent health care team* (pp. 72–73). Dorrance Publishing.

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.

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

Market Data Retrieval. (2001). A D&B Company: Shelton, CT.

Nagy, W. E., & Herman, P. E. (1987). Breadth and depth of vocabulary knowledge: implications for acquisition and instruction. In. M. G. McKeown & M. E. Curtis (Eds). *The Nature of Vocabulary Acquisition.* Princeton, NJ: Lawrence Erlbaum Associates.

National Early Literacy Panel. (2008). Developing Early Literacy: Report of the National Early Literacy Panel. http://lincs.ed.gov/publications/pdf /NELPReport09.pdf

National Reading Panel. (2000). *Teaching children to read: An Evidence-Based assessment of the scientific research literature on reading and its implications for reading instruction.* Bethesda, MD: National Reading Panel.

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.

Renaissance Learning. (2005). *Correlation between Michigan Literacy Progress Profile and Star Early Literacy.* Wisconsin Rapids, WI: Author.

Snow, C. E., Burns, M. E. & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Spearman, C. (1904) "The Proof and Measurement of Association between Two Things". *The American Journal of Psychology,* 15 (1), 72–101 JSTOR 1412159.

Texas Education Agency and the University of Texas System. *TPRI 2004–2006.* Texas Education Agency and the University of Texas System.

Trelease, J. (1995). *The read-aloud handbook.* New York: Penguin Books.

University of Oregon Center on Teaching and Learning. *DIBELS: Dynamic Estimators of Basic Early Literacy Skills.* Eugene, OR: University of Oregon.

Williams, K. T. (2001). *GRADE: Group Reading Assessment and Diagnostic Evaluation.* Circle Pines, MN: American Guidance Service.

Index

A

Access levels, 8
Adaptive Branching, 3, 5
Age and school grade, relationship to Star Early Literacy scores, 49
Alphabetic Principle, 10, 12, 26, 76

B

Bayesian-modal item response theory (IRT), 37
    estimation method, 90
Blueprint domains, 10, 24
    Comprehension Strategies and Constructing Meaning, 10, 17
    Numbers and Operations, 10, 18
    Word Knowledge and Skills, 10, 12
Blueprint skills, 24
Blueprint skill sets, 24
Blueprint sub-domain prescriptions, 12
Blueprint sub-domains, 10, 24, 76
    Alphabetic Principle, 10, 12, 26
    Concept of Word, 10, 13, 26
    Early Numeracy, 10, 18
    Paragraph-Level Comprehension, 10, 17, 26
    Phonemic Awareness, 10, 13, 26
    Phonics, 10, 14, 26
    Sentence-Level Comprehension, 10, 17, 26
    Structural Analysis, 10, 16, 26
    Visual Discrimination, 10, 13, 26
    Vocabulary, 10, 17, 26

C

Calibration
    initial items, 28
    item and scale, 28
    sample description, 29
Capabilities, 8
Center on Response to Intervention, 58
Chi-square test, 78
Common Core State Standards, 10
Comparative fit index (CFI), 78
Comprehension Strategies and Constructing Meaning, 10, 17, 76
Concept of Word, 10, 13, 26, 76
Conditional standard error of measurement (CSEM), 37, 41, 42, 45
Confirmatory Factor Analyses (CFA), 78
Content development, 10
Content specification, 10
Content structure, 24
Content validity, 48
Conversion tables, 99
Criterion measures for concurrent and predictive validity analyses, 55
Criterion-referenced scores, 90
Cronbach's alpha, 43
CRTI, 58

D

Data encryption, 8
Description of the program, 1
DIBELS, 63, 65
DIBELS Oral Reading Fluency (DORF), 69
Dynamic calibration, 28, 37

E

Early Numeracy, 10, 18, 76
Emergent Readers, 92, 104, 105
Empirical item response functions (EIRF), 33
Estimated Oral Reading Fluency (Est. ORF), 93
    concurrent validity, 69
Evidence of technical adequacy for informing screening and progress monitoring decisions, 57

Exercise questions, 5
Exploratory factor analysis (EFA), 77

G

Generic reliability, 42
Gifted and talented (G&T), 85
Goodness of fit index (GFI), 78
GRADE, 63, 65
Grade Equivalent (GE), 95
Growth norms, 81, 88

H

Hands-on exercise, 5

I

Individualized tests, 8
Instructions, repeating, 7
Item bank, 12
Item calibration, 28
Item development, 10, 22, 23
    content structure, 24
    metadata requirements and goals, 25
    readability guidelines, 26
    tagging for "Requires Reading", 24
    text, 26
Item response function, 32
Item Response Theory (IRT), 41, 76
    calibration, 28
Item retention, 34

K

Kuder-Richardson Formula 20 (KR-20), 43

L

Length of test, 5
Lexile® Measures, 98
Lexile® Ranges, 98
Literacy classification, 92

M

Market Data Retrieval (MDR), 29, 83
Maximum-Likelihood IRT estimation procedure, 90
Measurement precision, 41
Meta-analysis of validation study validity data, 56
Metadata requirements and goals, 25
Michigan Literacy Progress Profile (MLPP), 63, 64
Multi-Tiered Systems of Support (MTSS), 57

N

National Center for Educational Statistics (NCES), 83
National Center on Intensive Intervention (NCII), 58, 60
Normal Curve Equivalent (NCE), 54, 94
Normed fit index (NFI), 78
Norming, 81
    2017 Star Early Literacy norms, 81
    background, 81
    data analysis, 86
    geographic region, 84
    growth norms, 81, 88
    previous Star Early Literacy norms, 81
    sample characteristics, 82
    sample size, 83
    school size, 84
    socioeconomic status, 84
    test administration, 86
    test score norms, 81
Norm-referenced scores, 90
Numbers and Operations, 10, 18, 76

O

Office of Civil Rights (OCR), 85
On-line data collection, 37
Other tests, relationship to Star Early Literacy scores, 53
Overview of the program, 1

P

Paragraph-Level Comprehension, 10, 17, 26, 76
Password entry, 9
Percentile Rank (PR), 93
Phonemic Awareness, 10, 13, 26, 76
Phonics, 10, 14, 26, 76
Practice session, 5
Predicting reading performance from Star Early Literacy scores, 67
Pretest instructions, 4
Probable Readers, 93, 104, 106
Program design, 3
    Adaptive Branching, 5
    hands-on exercise, 5
    practice session, 5
    pretest instructions, 4
    test interface, 4
    test length, 5
    test repetition, 6
    time limits, 7
Program overview, 1
Progress monitoring, 59
    additional research, 62
    characteristics of assessments, 60
PROPORTION, 77

R

Rasch 1-parameter logistic IRT mode, 76
Rasch 1-parameter logistic response model, 90
Rasch ability estimates, 71
Rasch ability scores, 71, 91
Rasch IRT model, 35
Rasch model, 32
Rasch scores, 81, 86
Rating instrument, 51
    items used, 51
Readability guidelines, 26
Readers
    Emergent, 92, 104, 105
    Probable, 93, 104, 106
    Transitional, 93, 104, 105
Relationship of Star Early Literacy scores
    to age and school grade, 49
Reliability, 41
    generic, 42
    retest (alternate forms), 44
    split-half, 44
Repeating instructions, 7
Response to Intervention (RTI), 57
Retest (alternate forms) reliability, 44
Root mean square error of approximation (RMSEA), 78
RTI Action Network, 58, 60
Running Record, 63

S

Scale calibration, 28
Scales
    Star Early Literacy scale scores, 81, 87
    Unified, 38, 41, 81, 82, 87, 91
Scale Scores, 71, 91
School grade and age, relationship to Star Early Literacy scores, 49
Scores
    criterion-referenced, 90
    definitions, 90
    Normal Curve Equivalent (NCE), 54
    norm-referenced, 90
    relationship to age and school grade, 49
    relationship to other tests, 53
    relationship to skills ratings, 50, 52
    Scaled Score, 91
    Star Early Literacy scale scores, 81, 87
Score scales, definition and development, 35
Scoring, 37
Screening, 58
    key categories, 59
Security, 7
Sentence-Level Comprehension, 10, 17, 26, 76
Skill set scores, 92
Skills ratings, relationship to Star Early Literacy scores, 50, 52
Spearman-Brown formula, 44
Split application model, 8
Split-half reliability, 44
Standard error of measurement (SEM), 41, 45

Standardized root mean square residual (SRMR), 78
Star Early Learning, 82
Star Early Literacy, compared with conventional tests, 96
Star Early Literacy scale scores, 81, 87
Star Reading, 32
Structural Analysis, 10, 16, 26, 76
Student Growth Percentile (SGP), 88, 96
Sub-domain scores, 92

T

Test blueprint characteristics, 11
Test interface, 4
Test item design guidelines
- answer options, 21
- graphics, 19
- language and pronunciation, 21
- phonics and phonemic awareness, 19
- screen layout, 18
- simplicity, 18
- text, 19

Test length, 5
Test monitoring, 9
Test organization, 24
Test repetition, 6
Test score norms, 81
Test security, 7
- access levels and capabilities, 8
- data encryption, 8
- individualized tests, 8
- split application model, 8
- test monitoring/password entry, 9

Time limits, 7
TPRI, 63, 65
Transitional Readers, 93, 104, 105

U

Unified scale, 38, 41, 81, 82, 87, 91

V

Validation study data, 53
Validity, 48
- current and original Star Early Literacy versions, 71
- early numeracy test items, 73
- meta-analysis of the validation study validity data, 56
- rating instrument, 51
- relationship of Star Early Literacy scores to age and school grade, 49
- summary, 79
- validation study data, 53

Visual Discrimination, 10, 13, 26, 76
Vocabulary, 10, 17, 26, 76

W

Word Knowledge and Skills, 10, 12, 76
Words read correctly within one minute (WCPM), 69

Z

Zone of Proximal Development (ZPD), 104, 105

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

R43276.191112