RENAISSANCE LEARNING™

STAR Early Literacy™ Technical Manual

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

8/2015 SELRP

Contents

Introduction

Overview

STAR Early Literacy Enterprise is a computer-adaptive assessment instrument designed to measure the early literacy skills of beginning readers. STAR Early Literacy Enterprise addresses the need to determine children's mastery of literacy concepts that are directly related to their future success as readers. STAR Early Literacy Enterprise 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 Enterprise 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 Enterprise 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 eleven minutes. Despite its brevity, STAR Early Literacy Enterprise 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 Enterprise is designed specifically for repeated administration throughout the school year. STAR Early Literacy

Enterprise 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 is designed to provide teachers with 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 Enterprise 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 Enterprise 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 Enterprise:

- 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

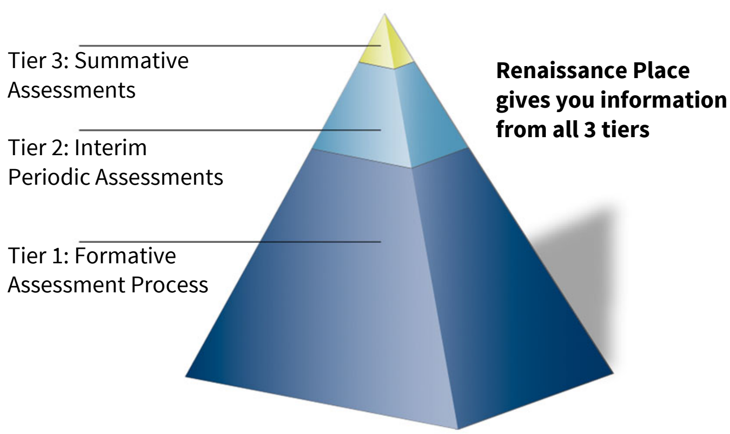
- timely and accurate information that can be used to plan literacy instruction and intervention
- a tool that enables them to capture a comprehensive picture of student literacy skills in ten sub-domains

- Helps teachers monitor student progress based on the specific literacy needs of each student.

Three Tiers of Student Information

The Renaissance Place edition of STAR Early Literacy Enterprise helps teachers accurately assess students' key early literacy skills in less than 15 minutes. This computer program also helps educators accelerate learning and increase motivation by providing immediate, individualized feedback on student academic tasks and classroom achievement. All key decision-makers throughout the district can easily access this information.

The Renaissance Place database stores all three levels of student information including the Tier 2 data from STAR Early Literacy Enterprise:



Tier 1: Formative Classroom Assessments

Formative classroom assessments provide daily, even hourly, feedback on students' task completion, performance, and time on task. Renaissance Learning Tier 1 programs include Accelerated Reader, MathFacts in a Flash, Accelerated Math, English in a Flash, and NEO/NEO 2.

Tier 2: Interim Periodic Assessments

Interim periodic assessments help educators match the level of instruction and materials to the ability of each student, measure growth throughout the year, predict outcomes on mandated state tests, and track growth in student achievement longitudinally, facilitating the kind of growth analysis recommended by state and federal organizations. Renaissance Learning Tier 2 programs include STAR Early Literacy Enterprise, STAR Math, STAR Math Enterprise, STAR Reading, and STAR Reading Enterprise.

Tier 3: Summative Assessments

Summative assessments provide quantitative and qualitative data in the form of high-stakes tests. The best way to ensure success on Tier 3 assessments is to monitor progress and adjust instructional methods and practice activities throughout the year using Tier 1 and Tier 2 assessments.

Design of STAR Early Literacy Enterprise

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

Another fundamental 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 Enterprise'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 Enterprise 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.

In April of 2013, the STAR Apps on iPad® was released, allowing students to take a STAR Early Literacy test on an iPad®. Students tap the answer of choice and then tap **Next** to enter the answer.

Pretest Instructions

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

Hands-On Exercise

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

1. The student's ability to move the mouse pointer to a target, and to click the mouse pointer on the target *or*

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.

Students must demonstrate proficiency in using the mouse or keyboard 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 Enterprise 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 Enterprise will halt the testing session and tell the student to ask the teacher for help.

Adaptive Branching/Test Length

STAR Early Literacy Enterprise's branching control uses a proprietary approach somewhat more complex than the simple Rasch maximum information Item Response Theory (IRT) model. The approach used in STAR Early Literacy Enterprise 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, STAR Early Literacy Enterprise begins the first administration of the test 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 Enterprise 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 Enterprise 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 Enterprise 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 content sub-domains assessed. The average length of time to complete a STAR Early Literacy Enterprise test (not including the pretest instructions) is 8.5 minutes, with a standard deviation of approximately 2 minutes. Most students will be able to complete a STAR Early Literacy Enterprise test in under 12 minutes, including pretest instructions, and almost all will be able to do so in less than 15 minutes.

Test Repetition

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

The STAR Early Literacy Enterprise item bank contains more than 2,300 items, so students can test often without getting the same questions more than once. STAR Early Literacy Enterprise 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 Enterprise 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 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 Enterprise 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 Enterprise software, when students log in, they do not have access to the same functions that teachers, administrators, and other personnel can access. Students are allowed to test, but they have no other tasks available in STAR Early Literacy Enterprise; 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 Enterprise 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 Enterprise data or access or change it with other software.

Access Levels and Capabilities

Each user's level of access to a Renaissance Place program depends on the primary position assigned to that user and the capabilities the user has been granted in Renaissance Place. 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 Place also allows you to restrict students' access to certain computers. This prevents students from taking STAR Early Literacy Enterprise tests from unauthorized computers (such as home computers). For more information on student access security, see the *Renaissance Place Software Manual*.

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 Enterprise 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 Place (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 Enterprise software.

Test Monitoring/Password Entry

Monitoring of student tests is another useful security feature of STAR Early Literacy Enterprise. 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.

While STAR Early Literacy Enterprise 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.

Psychometric Characteristics

The following sections provide an overview of the content of the STAR Early Literacy Enterprise test, its length in both number of items and administration time, and also its Adaptive Branching feature, the test scores it yields, and how those scores are distributed.

Content

Every STAR Early Literacy Enterprise 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 sub-domain are administered in every test. A summary of the test blueprint for STAR Early Literacy Enterprise 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 sub-domain.

Each STAR Early Literacy Enterprise 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 skill set.

Sub-Domain Prescriptions

For the first test a student takes during a school year, the number of items administered from each 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. These scale score bands define 5 literacy classifications, set out below in Table 1.

Table 1 lists the number of items from each STAR Early Literacy Enterprise sub-domain to be administered by grade or literacy classification to students in each of grades pre-K through 3. Students in grades higher than 3 are subject to the grade 3 prescriptions.

Table 1: STAR Early Literacy Enterprise: Numbers of Items per Domain and Sub-Domain to Be Administered at Each Grade on the First Test of the School Year (and at Each of 5 Literacy Classifications at Other Times)

	Pre-K	Grade K	Grade 1	Grade 2	Grade 3
	Literacy Classification				
	Emergent Reader		Transitional Reader		Probable Reader 775–900
Domains and Sub-Domains	Early 300–487	Late 488–674	Early 675–724	Late 725–774	
Word Knowledge and Skills Domain					
Alphabetic Principle	6	5	3	1	1
Concept of Word	5	4	2	1	1
Visual Discrimination	5	5	2	1	1
Phonemic Awareness	5	4	4	3	1
Phonics	0	3	5	4	3
Structural Analysis	0	0	0	3	4
Vocabulary	1	1	3	3	3
Comprehension Strategies and Constructing Meaning Domain					
Sentence-Level Comprehension	0	0	3	3	4
Paragraph-Level Comprehension	0	0	0	3	4
Numbers and Operations Domain					
Early Numeracy	5	5	5	5	5
Total	27	27	27	27	27

Additionally, restrictions in the software program ensure that questions that require the ability to read are not administered to students below the first grade. "Content and Item Development" on page 15 contains a detailed list of the ten literacy sub-domains and the 41 skill sets assessed by STAR Early Literacy Enterprise.

Test Length

Each STAR Early Literacy Enterprise session administers 27 test items tailored to the age, grade placement, and actual performance level of the student. Test length is consistent with the average attention span of young children. Dukette and Cornish "estimate for sustained attention to a freely chosen task range from about five minutes for a two-year old child, to a maximum of around 20 minutes in older children and adults." Others feel there is a formula for calculating the attention span using the child's age. Although the expert opinion varies on what the average student's attention span is, it is still apparent when working with younger children that their attention span is shorter than that of adults or older children and often falls into the range of 5–20 minutes.

Test Administration Time

A STAR Early Literacy Enterprise test typically takes less than 15 minutes to administer, including optional pre-test instructions and mouse training. During research and development, pre-test instructions and mouse training were not used; about 84 percent of all students completed the test in 10 minutes or less.

Adaptive Branching

STAR Early Literacy Enterprise selects items one at a time, based on a continually updated estimate of the student's ability level. Initially, this estimate is based on the student's age and grade placement. Subsequently, it is based on the student's actual performance on previous tests and during the current one. Using Adaptive Branching, the software chooses test items on the basis of content and difficulty, with the objective of matching item difficulty to the student's ability, and producing an average of 75 percent correct. This Adaptive Branching process is based on the branch of psychometrics called item response theory (IRT).

Score Scales

STAR Early Literacy Enterprise reports three different kinds of scores: Scaled Scores, Sub-domain Scores, and Skill Set Scores. Scaled Scores provide a global measure of the student's current ability. They are derived directly from the updated ability estimate computed after the last test question. Sub-domain Scores are separate estimates of the student's proficiency, expressed on a percent mastery scale, in each of the ten literacy sub-domains. Like Sub-domain Scores, Skill Set Scores are percent mastery estimates, but they are reported for each of the 41 STAR Early Literacy Enterprise skill sets.

Some reports (Growth, Screening, Summary, and Diagnostic–Student [also called the Student Diagnostic Report Skill Set Scores]) also include Estimated Oral

Reading Fluency (Est. ORF) Scores, which estimate a student's ability to read words quickly and accurately. (See page 141 for more information on Scaled Scores, Sub-domain Scores, Skill Set Scores, and Estimated Oral Reading Fluency Scores.)

STAR Early Literacy Enterprise and the Common Core State Standards

The Common Core State Standards (CCSS) that are directed toward fostering students' most basic understanding and working knowledge of concepts of print, the alphabetic principle, and other basic conventions of the English writing system are called Foundational Skills for Kindergarten through Grade 5. These early literacy skills are divided into four areas:

- Print Concepts
- Phonological Awareness
- Phonics and Word Recognition
- Fluency

In the CCSS, each level from kindergarten through grade 5 has grade-specific standards based in the anchor expectations for that grade level. The grade-by-grade expectations delineate steady growth in skills acquisition, resulting in increasingly sophisticated understanding.

STAR Early Literacy Enterprise is a pre-kindergarten through grade 3 assessment. The assessment focuses on measuring student performance in the following:

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

Research shows that measures in alphabetic principle, visual discrimination, phonemic awareness, phonological awareness, vocabulary, and early decoding are correlated with, and in some instances predictive of, later literacy achievement. Thus, STAR Early Literacy Enterprise provides valuable information regarding the development of early literacy skills.

The bulk of the skills assessed in STAR Early Literacy Enterprise are also found in the CCSS Foundational Skills. In addition, STAR Early Literacy Enterprise assesses some skills in three other areas included in the CCSS broader standards: the Reading Standards for Literature K-5, the Reading Standards for Informational Text K-5, and the skills for Vocabulary Acquisition and Use found in the Language Standards K-5. These sets of standards focus on key ideas and details, craft and structure, integration of knowledge, and vocabulary.

STAR Early Literacy has grade-level anchors and specific grade-level expectations largely based on the same standards as CCSS.

Resources consulted to determine the set of skills most appropriate for assessing reading development include the following:

Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects. 2010. http://www.corestandards.org/the-standards

Early Childhood Learning and Knowledge Center. *The Head Start Leaders Guide to Positive Child Outcomes.* September, 2003. http://eclkc.ohs.acf.hhs.gov/hslc /hs/resources/ECLKC_Bookstore/PDFs/HeadStartGuidePositiveChild Outcomes.pdf

National Institute for Literacy, and National Center for Family Literacy. *Developing Early Literacy: Report of the National Early Literacy Panel.* 2008. http://www.nichd.nih.gov/publications/pubs/upload/NELPReport09.pdf

Report of the National Reading Panel: Teaching Children to Read. *Findings and Determinations of the National Reading Panel by Topic Area.* Last updated 08/31/2006. http://www.nichd.nih.gov/publications/nrp/findings.cfm

SEDL. *The Cognitive Foundations of Learning to Read: A Framework.* 2009. http://www.sedl.org/reading/framework/

Shanker, James L., and Eldon E. Ekwall. *Locating and Correcting Reading Difficulties. 8th Ed.* New Jersey: Merrill Prentice Hall. 2003

U.S. Department of Education, and Institute of Education Sciences (IES): National Center for Education Evaluation and Regional Assistance. *Assisting Students Struggling with Reading: Response to Intervention (RtI) and Multi-Tier Intervention in the Primary Grades.* February, 2009. http://ies.ed.gov/ncee/wwc/pdf/practiceguides/rti_reading_pg_021809.pdf

Content and Item Development

Content Specification

STAR Early Literacy Enterprise consists of 2,120 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.

Since the initial 2001 release of STAR Early Literacy 1.0, it has been a 25-item adaptive test of early literacy skills. From 2001 until the development of the Enterprise version, STAR Early Literacy has included test items measuring 7 literacy domains and 41 subordinate preliteracy and early literacy skill sets; STAR Early Literacy Enterprise content is organized into 3 domains, 10 sub-domains, and 41 skill sets.

The content of the STAR Early Literacy Enterprise item bank is based in part on extensive analysis of existing curricula and standards. The content of the STAR Early Literacy Enterprise item bank is aligned to and substantially influenced by the recently developed Common Core State Standards.

Items are organized into sub-domains that are similar to widely accepted early literacy standards. The content of STAR Early Literacy Enterprise is organized into 3 broad domains and 10 sub-domains as follows:

Domains:

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

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 Enterprise 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.

The categorization of skills into skill sets and domains in STAR Early Literacy Enterprise is 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 Enterprise 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 Enterprise 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 Enterprise. Furthermore, the assessment estimates students' oral reading fluency and displays these estimates on certain reports. (See page 111 for information on the Estimated Oral Reading Fluency scores.)

The STAR Early Literacy Enterprise Item Bank

Within each of the three STAR Early Literacy Enterprise domains, closely related skill sets are organized into sub-domains. The resulting hierarchical structure is domain, sub-domain, skill set, and specific skill. Tables 2–4 display the domains, sub-domains, skill sets, and skills.

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

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Alphabetic Principle	Alphabetic Knowledge	AP02A	Recognize lowercase letters	Pre-K
		AP02B	Recognize uppercase letters	Pre-K
		AP02C	Match lowercase with uppercase letters	K
		AP02D	Match uppercase with lowercase letters	K
		AP02E	Distinguish numbers from letters	K
	Alphabetic Sequence	AP03A	Identify the letter that comes next	K
		AP03B	Identify the letter that comes before	K
	Letter Sounds	AP04A	Recognize sounds of lowercase letters	K
		AP04B	Recognize sounds of uppercase letters	K

Table 2: Hierarchical Structure of the STAR Early Literacy Enterprise Item Bank: Word Knowledge and Skills Domain (Continued)

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Concept of Word	Print Concepts: Word Length	CW01A	Identify longest word	K
		CW01B	Identify shortest word	K
	Print Concepts: Word Borders	CW02A	Identify number of words (2–3)	K
	Print Concepts: Letters and Words	CW03A	Differentiate words from letters	K
		CW03B	Differentiate letters from words	K
Visual Discrimination	Letters	VS01A	Differentiate lowercase letters	Pre-K
		VS01B	Differentiate uppercase letters	Pre-K
		VS01C	Differentiate lowercase letters in mixed set	Pre-K
		VS01D	Differentiate uppercase letters in mixed set	Pre-K
	Identification and Word Matching	VS03A	Identify words that are different	K
		VS03B	Match words that are the same	K
		VS03C	Identify words that are different from a prompt	K
Phonemic Awareness	Rhyming and Word Families	PA01A	Match sounds within word families (named pictures)	Pre-K
		PA01B	Match sounds within word families (unnamed pictures)	Pre-K
		PA01C	Identify rhyming words (named pictures)	K
		PA01D	Identify nonrhyming words (named pictures)	K
	Blending Word Parts	PA02A	Blend onsets and rimes	K
		PA02B	Blend 2-syllable words	K
		PA02C	Blend 3-syllable words	K
	Blending Phonemes	PA03A	Blend phonemes in (VC) or (CVC) words	K
		PA03B	Blend phonemes in single-syllable words	1

Table 2: Hierarchical Structure of the STAR Early Literacy Enterprise Item Bank: Word Knowledge and Skills Domain (Continued)

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Phonemic Awareness (continued)	Initial and Final Phonemes	PA04A	Determine which word (picture) has an initial phoneme different from a prompt	K
		PA04B	Determine which word (picture) has a different initial phoneme	K
		PA04C	Match initial phoneme to a prompt (pictures)	K
		PA04D	Recognize same final sounds (pictures)	K
		PA04E	Determine which word (picture) has a final phoneme different from a prompt	K
		PA04F	Determine which word (picture) has a different final phoneme	K
	Consonant Blends (PA)	PA07A	Match consonant blend sounds (pictures)	K
	Medial Phoneme Discrimination	PA08A	Identify short vowel sounds (pictures)	K
		PA08B	Identify and match medial sounds (pictures)	K
		PA08C	Distinguish short vowel sounds (pictures)	K
		PA08D	Match long vowel sounds (pictures)	1
		PA08E	Distinguish long vowel sounds (pictures)	1
	Phoneme Segmentation	PA09A	Segment syllables in multisyllable words	K
		PA09B	Segment phonemes in single-syllable words	1
	Phoneme Isolation/Manipulation	PA10A	Substitute initial consonant (named pictures)	K
		PA10B	Substitute initial consonant (unnamed pictures)	K
		PA10C	Determine missing phoneme, initial or final	1
		PA10D	Substitute initial consonant in a prompt (pictures)	1
		PA10E	Substitute final consonant sound in a prompt (unnamed pictures)	1
		PA10F	Substitute final consonant (named pictures)	1
		PA10G	Substitute final consonant sound (unnamed pictures)	1
		PA10H	Substitute vowel sounds (pictures)	1

Table 2: Hierarchical Structure of the STAR Early Literacy Enterprise Item Bank: Word Knowledge and Skills Domain (Continued)

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Phonics	Short Vowel Sounds	PH02A	Match short vowel sounds (words)	1
		PH02B	Match short vowel sounds to letters	K
		PH02C	Decode CVC words	K
		PH02D	Recognize short vowel sounds (words)	1
		PH02E	Distinguish short vowel sounds (words)	1
		PH02F	Decode grade-appropriate words	1
	Initial Consonant Sounds	PH03A	Identify initial consonant sound (words)	K
		PH03B	Identify letter for initial consonant sound (words and letters)	K
	Final Consonant Sounds	PH04A	Match word to a given final consonant sound	1
		PH04B	Identify letter for a final consonant sound	1
	Long Vowel Sounds	PH01A	Identify long vowel sounds (words)	1
		PH01B	Match long vowel sounds to a prompt (words)	1
		PH01C	Distinguish long vowel sounds (words)	1
		PH01D	Match long vowel sounds to letters	1
		PH01E	Decode and recognize associated spelling patterns with long vowels (C-V-C-e)	1
		PH01F	Decode and recognize associated spelling patterns with long vowel open syllables	1
		PH01G	Decode and recognize associated spelling patterns with long vowel digraphs (including *y* as a vowel)	2
	Variant Vowel Sounds	PH14A	Identify variant vowel sounds	2
		PH14B	Identify variant vowel sounds (words)	2
		PH14C	Decode words with variant vowels and recognize associated spelling patterns	2

Table 2: Hierarchical Structure of the STAR Early Literacy Enterprise Item Bank: Word Knowledge and Skills Domain (Continued)

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Phonics (continued)	Consonant Blends (PH)	PH10A	Recognize initial consonant blends (words)	1
		PH10B	Distinguish consonant blends (words)	1
		PH10C	Recognize word with a consonant blend in a contextual sentence	1
		PH10D	Recognize associated spelling patterns of initial consonant blends	2
		PH10E	Recognize associated spelling patterns of final consonant blends	2
	Consonant Digraphs	PH12A	Identify a consonant digraph in a named word	1
		PH12B	Identify a consonant digraph in an unnamed word	1
		PH12C	Identify a contextual word containing a consonant digraph	1
		PH12D	Identify correct spelling of consonant digraphs in words	1
	Other Vowel Sounds	PH15A	Identify diphthong sounds in words	2
		PH15B	Decode words with diphthongs and recognize associated spelling patterns	2
		PH15C	Identify *r-controlled* vowel sounds in named and unnamed words	2
		PH15D	Decode words with *r*-controlled vowels and recognize associated spelling patterns	2
	Sound-Symbol Correspondence: Consonants	PH05A	Substitute initial consonants (words)	1
		PH05B	Substitute final consonants (words)	1
		PH05C	Substitute final consonant sound (named words)	1
		PH05D	Substitute final consonant sound (unnamed words)	1
	Word Building	PH19A	Identify words made by adding an initial consonant (unnamed words)	1
		PH19B	Identify words made by adding an additional medial letter (unnamed words)	1
		PH19C	Identify words made by adding an additional final letter (unnamed words)	1
		PH19D	Identify words built by adding one letter to an audio prompt	1

Table 2: Hierarchical Structure of the STAR Early Literacy Enterprise Item Bank: Word Knowledge and Skills Domain (Continued)

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Phonics (continued)	Sound-Symbol Correspondence: Vowels	PH06A	Substitute vowel sounds (words)	1
	Word Families/Rhyming	PH09A	Identify rhyming words (words)	1
		PH09B	Identify nonrhyming words (words)	1
		PH09C	Identify rhyming words (unnamed answers)	1
		PH09D	Identify rhyming words (unnamed prompt and answers)	1
		PH09E	Identify nonrhyming words (unnamed prompt and answers)	1
		PH09F	Identify onset/rime in named words	1
		PH09G	Identify onset/rime in unnamed words	1
		PH09H	Identify sounds within word families (named words)	1
		PH09I	Identify sounds within word families (unnamed words)	1
StructuralAnalysis	Words with Affixes	SA01A	Use knowledge of common affixes to decode words	2
	Syllabification	SA02A	Use knowledge of syllable patterns to decode words	2
		SA02B	Decode multisyllable words	3
	Compound Words	SA03A	Identify compound words (named words)	1
		SA03B	Identify words that are not compounds (named words)	1
		SA03C	Identify compound words (unnamed words)	1
		SA03D	Identify words that are not compounds (unnamed words)	1
		SA03E	Identify correctly formed compounds	2

Table 2: Hierarchical Structure of the STAR Early Literacy Enterprise Item Bank: Word Knowledge and Skills Domain (Continued)

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Vocabulary	Word Facility	VO01A	Match words to pictures	K
		VO01B	Read high-frequency words by sight	K
		VO01C	Identify new meanings for common multi-meaning words	K
		VO01D	Determine categorical relationships	K
		VO01E	Understand position words	K
		VO01F	Read grade-level sight words	1
		VO01G	Understand multi-meaning words	1
	Synonyms	VO02A	Identify synonyms of grade-appropriate words	1
		VO02B	Match words with their synonyms (words)	1
		VO02C	Identify synonym of a grade-appropriate word in a contextual sentence	1
		VO02D	Match words with their synonyms in paragraph context (assisted)	1
		VO02E	Match words with their synonyms in paragraph context (unassisted)	2
	Antonyms	VO03A	Identify antonyms of words	1
		VO03B	Identify antonyms of words in context (assisted)	1
		VO03C	Identify antonyms of words in context (unassisted)	2

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

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Sentence-Level Comprehension	Comprehension at the Sentence Level	SC01A	Listen and identify word in context	1
		SC01B	Read and identify word in context	1
Paragraph-Level Comprehension	Comprehension of Paragraphs	PC01A	Identify the main topic of a text	K
		PC01B	Listen to text and answer literal *who, what* questions	1
		PC01C	Listen to text and answer *where, when, why* questions	1
		PC01D	Read text and answer literal *who, what* questions	1
		PC01E	Read text and answer *where, when, why* questions	2

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

Sub-Domain	Skill Set	Skill ID	Skill	Grade
Early Numeracy	Number Naming and Number Identification	EN01A	Recognize numbers 0–20	K
	Number Object Correspondence	EN04A	Count 1–20	K
		EN04B	Recognize ordinal numbers 1st–10th	K
		EN04C	Compare sets of up to 5 objects	K
		EN04D	Identify the number of 10s in 10, 20, 30, 40, 50, 60, 70, 80, 90	1
	Sequence Completion	EN03A	Complete a picture pattern	K
		EN03B	Complete a sequence of numbers between 0 and 10 in ascending order	K
	Composing and Decomposing	EN05A	Add 1 to a set	K
		EN05B	Subtract 1 from a set	K
		EN05C	Add numbers with a sum up to 10 (pictures)	K
		EN05D	Subtract numbers with a minuend up to 10 (pictures)	K
	Measurement	EN02A	Compare sizes, weights, and volumes	K

Item Design Guidelines

Each of the items was written to the following specifications:

Simplicity

Items should directly address the sub-domain and skill set objective 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 should consist of some 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, and sound.

Text

For letter and word identification items, the type size should be relatively large, becoming smaller through the grades. The type size should be tied to items, so that it varies according to the developmental level of a student; in other words, easier items should have larger type than more difficult items because the difficulty will correspond roughly to grade placement.

All STAR Early Literacy Enterprise test items will be administered aurally by the computer, so there should be no need for printed directions on-screen. For any items that require on-screen directions, the type should be a serif font of appropriate size.

Every effort should be made to use common words as the target and distracter words in test items.

For phonemic awareness and phonics items, the 44 phonemes (speech sounds) that make up the English language should be 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.

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.

Answer Options

As a general rule, items should have three answer choices. Only one of the choices should be the correct answer. Answer choices should be arranged horizontally. For internal purposes the answers may be labeled A, B, and C, moving from left to right.

Distracters should be 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 a skill set, 100 out of 1,000 items in a sub-domain, and 300 out of 3,000 items in the item bank.

Language and Pronunciation

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 Enterprise

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 Enterprise assessment consists of 27 items selected adaptively from a bank of 2,120 multiple-choice items (as of January 2012, with hundreds of others in the field gathering calibration data of additional items during the spring of 2012) that tap knowledge and skills from 9 early literacy sub-domains and 1 named Early Numeracy. Each item contains two distracters. The STAR Early Literacy Enterprise 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 Enterprise test, a content-balancing blueprint ensures that a prescribed number of items from each sub-domain are administered. The number of items per sub-domain varies by grade and by score ranges on previously taken tests.

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

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

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

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

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 Enterprise 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 Enterprise, 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. To compensate for a lack of diversity, content featuring diversity will be emphasized in 2012.
- **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 Enterprise 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.

Each of the items was developed according to the following specifications:

- STAR Early Literacy Enterprise items are designed to efficiently assess targeted skills. Items should not take overly long for a student when taking a test. The items should be engaging so that students will enjoy reading them.

- Items directly address the sub-domain and skill set objective in the most straightforward manner possible.
- Each item consists 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, and sound.
- For phonemic awareness and phonics items, the 44 phonemes that make up the English language are used. Phonemes are depicted by one or more letters enclosed in forward slashes as shown in Table 5.
- Items have three answer choices. Only one is the correct answer. Answer choices are arranged horizontally. Answer choices are reasonable, but not tricky.
- Words and artwork used in answer choices are 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 a skill set, 100 out of 1,000 items in a sub-domain, and 300 out of the 3,000 items in the item bank.
- Language should be used consistently throughout the assessment. Standard protocols have been established for item administration that reflects consistent instructions. For example, if an item stem is repeated twice, it should be used for all items of the same type. One exception is 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. Some anomalies were introduced when reorganizing and combining a few of the skills, but efforts are underway to minimize the impacts.
- In items for Paragraph-Level Comprehension (PC), the use of first person is to be avoided in passages. This is done to preclude the awkward reference to the narrator as "this person" in asking the question at the end of the passage as is done here: Why is this person tired?
- Target or prompt words in vowel-sound items that contain *l* or *w* following a vowel are not used unless the skill addresses a variant vowel sound. (The consonants *l* and *w* somewhat distort the vowel sounds.) Some instances of this occur in the operational content. Every effort will be made not to repeat the practice.
- Efforts were made to avoid the confusion of s/z sounds as is demonstrated here:

  Look at the pictures: dress, pens, church. Pick the picture whose ending sound is different from likes...likes.
- Developers strive to create items that will work well in both the US and the UK. A list of words not to be used in items has been developed. For example, the

US word “flashlight” is referred to as “torch” in the UK, so developers will avoid using “flashlight” or “torch” in items.

- A decision was made to replace all offensive graphics: kill, kiss, etc. However, in some instances the word graphics were allowed. The thinking is that the sound of the word is less offensive when the concentration is on its sound rather than on its meaning.
- Only healthy foods will be represented in current and future development.
- All new items moving forward should contain either gender-neutral or cross-stereotypical situations. Girls will be shown participating fully in life and boys will be shown studying or reading or entertaining or caring for younger children.

Readability Guidelines

ATOS GLE and word counts will be 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 Enterprise item. Words used in scripts should be appropriate for the intended grade.

The content in STAR Early Literacy Enterprise 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 service edition of STAR Early Literacy and applies to STAR Early Literacy Enterprise):

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.

Text of Scripts/Audio Instructions

Blending parts of 2- or 3- syllable words (as in PA03) or Phoneme Segmentation (as in PA09) or Syllabification (SA02):

1. Each word will *first* be pronounced as shown in the pronunciation guide in the dictionary.

2. For the purpose of demonstrating blending of syllables or syllable segmentation, words will be pronounced using the word breaks given in the dictionary, not the pronunciation guides.

3. The sounds themselves will be pronounced as closely as is possible to the way the sounds are heard in the word.

    - Retain the accent(s) of the word (not all syllables will be equally accented).
    - A vowel in an unaccented syllable makes the schwa sound. Use that pronunciation in saying the whole word. However, when saying the sound within the syllable, use something a little closer to the short vowel sound (using Ehri's correctionist theory of word learning, 1998). A separate document on the research on syllabification is being prepared as of this writing.
    - The audio should ensure that the sounds are correctly pronounced and not turned into nonexistent syllables (not *muh* but *mmm*).

4. In *blending* the sounds in the word ("mmmmmmmaaaaaaaaannnnnnnn"). Do not stop between the sounds. Make certain that the sounds are not distorted as you stretch them out. Hold each sound long enough for the students to hear it individually. Stop sounds cannot be prolonged without distortion. When pronouncing words that begin with stop sounds (such as *t, k,* and *p*), pronounce the initial sound quickly and do not stretch it out. Clip the sound of a consonant stop sound at the end of a word.

5. In *segmenting* the syllables in a word, stop between the syllables.
6. The audio producers should ensure that the phonemes are correctly pronounced and not turned into nonexistent syllables (not *muh* but *mmm*).
7. “Pick the” is the preferred wording in the last sentence in STAR Early Literacy Enterprise scripts.

Core Progress Learning Progression for Reading and the Common Core State Standards

The Reading Foundational Skills (K-5) in CCSS are directed toward fostering students' understanding and working knowledge of concepts of print, the alphabetic principle, and other basic conventions of the English writing system. These foundational skills are necessary and important components of an effective, comprehensive reading program designed to develop proficient readers with the capacity to comprehend texts across a range of types and disciplines. The CCSS standards provide grade-level specific standards that delineate the progress toward these goals.

Much like in the CCSS, the Core Progress for Reading contains the foundational skills needed for learning to read. The Core Progress for Reading is a research-based and empirically supported learning progression for reading. It identifies the continuum of reading skills, strategies, and behaviors needed for students to be accomplished and capable readers. The continuum begins with emergent reading and progresses to the level of reading ability required for college and careers. The skills assessed in STAR Early Literacy Enterprise are a subset of this larger continuum of skills. STAR Early Literacy Enterprise assessment results are correlated to the Core Progress learning progression for reading.

Table 7 provides information regarding the correlation, or alignment, of STAR Early Literacy Enterprise skills with the CCSS Foundational Skills, including several vocabulary acquisition skills from the Language Standards.

Table 7: STAR Early Literacy Enterprise and CCSS Reading Standards Foundational Skills + Vocabulary Acquisition Skills from Standards Language Standards

	Kindergarten	Grade 1	Grade 2	Grade 3
Number of Foundational Skills per grade level	14	15	9	7
Number of Foundational Skills with an alignment to STAR Early Literacy Enterprise	14	15	9	7
Percentage of Foundational Skills aligned to STAR Early Literacy Enterprise skills	100%	100%	100%	100%

In addition to the number of CCSS Foundational and Vocabulary Skills in table 9, STAR Early Literacy Enterprise includes a total of 18 additional skills from the Reading Standards for Literature and Informational Text in Kindergarten, Grade 1, and Grade 2.

Psychometric Research Supporting STAR Early Literacy Enterprise

Since 2000, there have been three major phases of research leading to the development and publication of STAR Early Literacy in 2001, and subsequently to the publication of STAR Early Literacy Enterprise in 2012. These are referred to as the 2000 Calibration Study, the 2001 Validation Study, and the 2012 STAR Early Literacy Enterprise Research Study.

The 2000 Calibration Study is described in detail in the section on Item Calibration below. The 2001 Validation Study is described in the Validity section, beginning on page 55. The 2012 STAR Early Literacy Enterprise Research Study is described following the Validation Study, beginning on page 86.

Item Calibration

Background

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. For the Calibration Study, 246 test forms were designed, and the 2,929 STAR Early Literacy items were distributed among these forms. Every form contained ten mouse training items, ten practice items, and forty test items (the keyboard was not used to enter answers for this study). The forms were graded as to developmental level:

Level A forms were designed for pre-kindergarten and kindergarten, Level B was designed for students in first grade, and Level C was designed for students in second and third grade.

Because all STAR Early Literacy test items include computerized graphics and audio, these calibration test forms were all computer-administered. Over 46,000 computer-administered calibration tests were given to a nationwide sample of students in pre-kindergarten through grade 3.

Over 300 schools in the United States participated in the Calibration Study. 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 8 compares certain sample characteristics of the students participating in the Calibration Study against national percentages of the same characteristics.

Table 8: 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 8, additional information about participating schools and students was collected. This information is summarized in Table 9, Table 10, and Table 11. These tables also include national figures based on 2001 data provided by MDR.

Table 9: 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 10: 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 11: 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 12 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 12: 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 15.

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 44.

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 55.

Statistical Analysis: Fitting the Rasch IRT Model to the Calibration Data

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. 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]

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.

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).

Selection of Items from the Calibration Item Bank

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 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 contains 2,350 items;[2] of these, 2,332 are available for use in adaptive testing and the other 18 are used as practice items.

2. The item count stood at 2,369 until it was reduced to 2,350 with the release of STAR Early Literacy RP version 3.3.

Dynamic Calibration

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 pupils.

Pupil 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.

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.[3] A system of integer numbers ranging from 300 to 900 was chosen as the score reporting scale for STAR Early Literacy. More information about the score scale is presented in "Score Definitions" on page 110.

3. 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.

Reliability and Measurement Precision

Reliability is a measure of the degree to which test scores are consistent 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 Enterprise, 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 Enterprise 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 average amount of measurement precision in a specific examinee group or population as a whole. In STAR Early Literacy Enterprise, the conditional standard error of measurement (CSEM) is an estimate of the unreliability of each individual test score. A reliability coefficient is a single value that applies 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 Enterprise test scores. Both generic reliability and split-half reliability are estimates of the internal consistency reliability of a test.

Generic Reliability

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

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

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.

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

where the summation is over the squared values of the reported CSEM for students i = 1 to n. In each STAR Early Literacy 3.x and higher 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 2.0 norming data resulted in the generic reliability estimates shown in the rightmost column of Table 13 on page 49. 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 versions prior to STAR Early Literacy Enterprise.

While generic reliability does provide a plausible estimate of measurement precision, it is a theoretical estimate, as opposed to traditional reliability coefficients, which are more firmly based on item response data. Traditional internal consistency reliability coefficients such as Cronbach's alpha and Kuder-Richardson Formula 20 (KR-20) cannot be calculated for adaptive tests. However, 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,[4] 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 below and more firmly grounded in the item response data.

The third column of Table 13 on page 49 contains split-half reliability estimates for STAR Early Literacy, calculated from the Validation Study data. Split-half scores were based on the first 24 items of the test; scores based on the odd- and the even-numbered items were calculated. The correlations between the two sets of scores were corrected to a length of 25 items, yielding the split-half reliability estimates displayed in Table 13 on page 49.

Test-Retest 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 reliability coefficient. This approach was used for STAR Early Literacy in both the Calibration Study and the Validation Study. In the Calibration Study, the participating schools were asked to administer two forms of the calibration tests, each on a different day, to a small fraction of the overall sample. This resulted in a test-retest reliability subsample of about 14,000 students who took different forms of the 40-item calibration test. In the Validation Study, the schools were asked to administer computer-adaptive STAR Early Literacy tests twice to every student. Over 90 percent of the Validation Study sample took two such tests over an interval of several days. From the two studies, we have two different sets of estimates of STAR Early Literacy retest reliability—one derived from two administrations of the 40-item non-adaptive Calibration Study tests, and one derived from two administrations of the 25-item adaptive Validation Study tests.

The retest reliability data from the Calibration Study provide an approximate measure of the reliability of tests constructed from items of the kind developed for use in the STAR Early Literacy item bank.

The retest reliability data from the Validation Study provide a more definitive measure of STAR Early Literacy reliability, because the tests were adaptively

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

administered, using only the items that were retained in the STAR Early Literacy item bank following the Calibration Study.

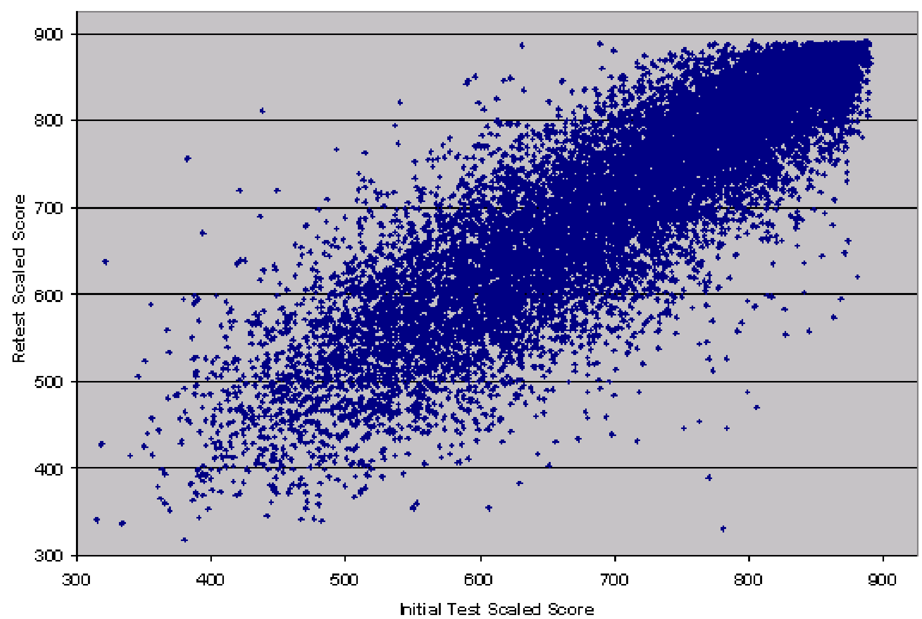
Retest reliability data from both studies are reported in the following sections.

Calibration Study Data

As mentioned earlier, the Calibration Study included a test-retest Reliability Study, in which selected students took calibration tests twice. The two tests were administered on different days, and each student took a different form on retest to minimize repetition of the same items. The correlation of students' scores on their first and second tests provides one measure of the reliability of STAR Early Literacy tests.

Over 14,000 students took part in the retest Reliability Study. Figure 1 shows a scatterplot of students' scores on initial test and retest. As the figure indicates, the correlation was substantial: 0.87 overall.

Figure 1: Scatterplot of STAR Early Literacy Initial and Retest Scaled Scores from the Calibration Study (N = 14,252 Students; Correlation = 0.87)



Validation Study Data

As in the Calibration Study, some students participating in the Validation Study took STAR Early Literacy on two occasions, separated by a few days. The items administered to a student during the initial test were not repeated during the second test.

The correlation of students' scores on their first and second tests provides a measure of the reliability of STAR Early Literacy tests that have been adaptively administered.

Over 9,000 students took part in the retest Reliability Study. Checks were made for valid test data on both test administrations and to remove cases of apparent motivational discrepancies. Figure 2 shows a scatterplot of 9,146 students' scores on initial test and retest. As the figure indicates, the correlation was 0.86 overall.

Figure 2: Scatterplot of STAR Early Literacy Initial and Retest Scaled Scores from the Validation Study (N = 9,146 Students; Correlation = 0.86)

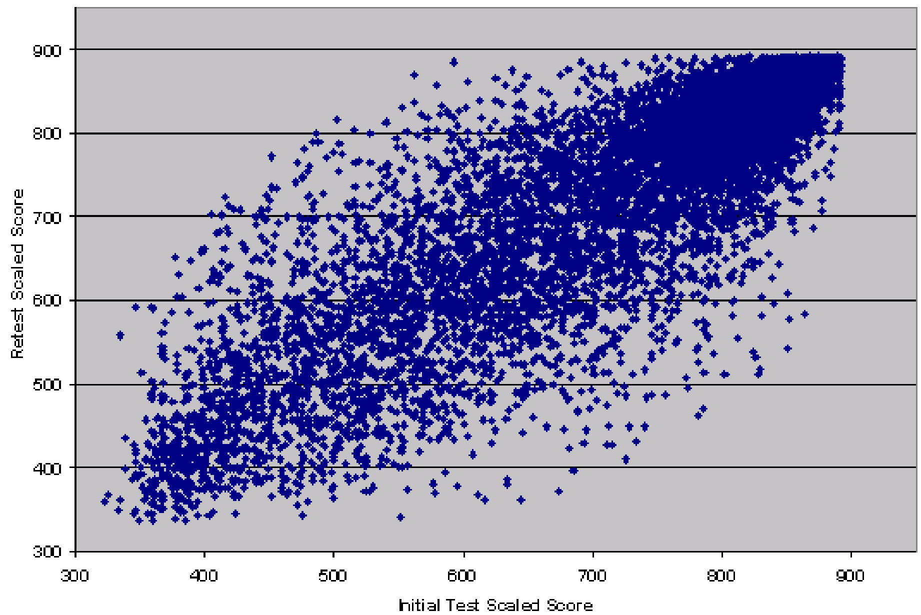


Table 13 lists the detailed results of the split-half, retest and generic reliability analyses of STAR Early Literacy Scaled Scores on versions prior to the Enterprise version, both overall and broken out by grade. Given the ways in which STAR Early Literacy will be used, the split-half reliability coefficient from the overall sample may be the most appropriate one for evaluating the psychometric characteristics of STAR Early Literacy as a measurement instrument. However, because there will be some interest in STAR Early Literacy's reliability at the school grade level, the table also includes grade-by-grade reliability data.

Table 13: Scaled Score Reliability Estimates by Grade for Pre-Enterprise Versions

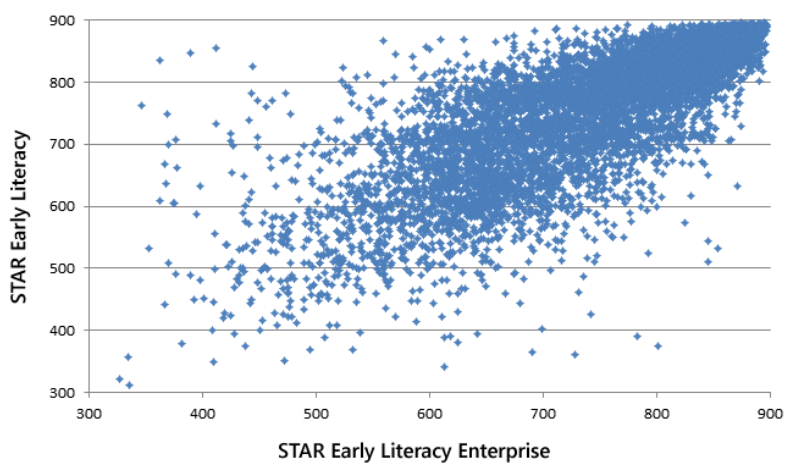
	N	Split-Half Reliability	Retest Reliability	Generic Reliability
All Grades	9,146	0.91	0.86	0.92
Pre-Kindergarten	529	0.84	0.63	0.85
Kindergarten	2,107	0.75	0.66	0.77
Grade 1	2,515	0.82	0.70	0.85
Grade 2	1,971	0.82	0.68	0.85
Grade 3	2,024	0.83	0.66	0.85

STAR Early Literacy Enterprise Equivalence Study Data

In February and March 2012, the first STAR Early Literacy Enterprise research study was conducted. In that study, more than 7,000 schoolchildren in grades K through 3 were administered two versions of STAR Early Literacy: the older, 25-item "service" version and the new, 27-item Enterprise version. In addition, teachers rated the children's early literacy skills using a 10-item skills inventory designed to align to key skills in the Common Core State Standards at each of the four grades.

Figure 3 shows a scatter plot of scores on the two tests for students of all grades combined; the correlation coefficient was 0.78.

Figure 3: Scatterplot of STAR Early Literacy Enterprise Research Study Scale Scores, STAR Early Literacy Service versus STAR Early Literacy Enterprise



The data displayed in the plot in Figure 3 were used to evaluate the reliability and some aspects of the validity of STAR Early Literacy. Data from the STAR Early Literacy Enterprise test were used to calculate the estimates of its internal consistency reliability using two different approaches: generic and split-half reliability estimation. Score data from the two STAR Early Literacy versions were used to calculate correlation coefficients between them—estimates of test-retest or, more correctly, alternate test reliability.

In April 2014, as part of the 2014 norming analyses, reliability analyses for STAR Early Literacy Enterprise were also conducted. These reliability analyses, which were based on all data from the 2012–2013 school year, computed generic internal consistency reliability, retest reliability and split half reliability. The split half reliability was computed by rescoring the randomly sampled test records for the odd and even items using a Rasch model analysis, converting the Rasch ability to Scaled Scores and then adjusting the half-length assessments (13 items) to full length assessments (27 items) using the Spearman Brown formula. The split half reliability presented in Table 14 is the correlation between the Scaled Scores estimated from the odd and even numbered items adjusted to a full assessment length of 27 items. This reliability analysis was conducted within and across grades and results are presented in Table 14. The table shows the grade, sample size N and reliability coefficients for generic internal consistency reliability, retest reliability, and split half reliability.

Table 14: Internal Consistency, Retest Reliability, and Split Half Reliability of STAR Early Literacy Enterprise (Assessments Taken in the 2012–2013 School Year)

	Internal Consistency Estimates				Retest Reliability	
Grade	N	Generic Reliability	N	Split-Half Reliability	N	Reliability Coefficient
Pre-K	48,078	0.80	2,500	0.81	3,517	0.57
K	48,078	0.79	2,500	0.78	3,517	0.49
1	48,078	0.81	2,500	0.79	3,517	0.49
2	48,078	0.84	2,500	0.85	3,517	0.60
3	48,078	0.89	2,500	0.90	3,517	0.73
All	240,390	0.90	12,500	0.90	17,585	0.78

Table 15 below lists the split-half reliability coefficients obtained from the analyses of each of the ten Sub-Domain scores. Sub-Domain Score reliability data are presented overall and by grade. Table 16 lists similar data for the Skill Set Scores. The split-half reliability estimates in Tables 15 and 16 are based on the STAR Early Literacy Enterprise Equivalence Study conducted in February and March of 2012.

Table 15: Sub-Domain Score Split-Half Reliability, Overall and by Grade

Sub-Domain Score	Grades				
	Overall	K	1	2	3
Alphabetic Principle	0.84	0.76	0.82	0.85	0.86
Concept of Word	0.84	0.77	0.83	0.85	0.87
Early Numeracy	0.84	0.76	0.82	0.84	0.86
Paragraph-Level Comprehension	0.85	0.74	0.81	0.83	0.82
Phonemic Awareness	0.85	0.74	0.81	0.83	0.83
Phonics	0.85	0.75	0.81	0.84	0.84
Sentence-Level Comprehension	0.85	0.74	0.81	0.83	0.84
Structural Analysis	0.85	0.74	0.81	0.83	0.83
Visual Discrimination	0.84	0.78	0.83	0.85	0.87
Vocabulary	0.85	0.75	0.82	0.84	0.84

Table 16: Skill Set Score Split-Half Reliability, Overall and by Grade

Skill Set Score	Grades				
	Overall	K	1	2	3
Alphabetic Principle Sub-Domain					
Alphabetic Knowledge	0.84	0.78	0.83	0.85	0.87
Alphabetic Sequence	0.84	0.77	0.82	0.85	0.87
Letter Sounds	0.84	0.74	0.82	0.84	0.85
Concept of Word Sub-Domain					
Print Concepts: Word Length	0.84	0.78	0.83	0.85	0.87
Print Concepts: Word Borders	0.84	0.75	0.82	0.84	0.86
Print Concepts: Letters and Words	0.84	0.79	0.83	0.85	0.88
Early Numeracy Sub-Domain					
Number Naming and Number Identification	0.84	0.77	0.82	0.84	0.86
Sequence Completion	0.84	0.74	0.82	0.84	0.86
Number Object Correspondence	0.84	0.75	0.82	0.84	0.86

Table 16: Skill Set Score Split-Half Reliability, Overall and by Grade (Continued)

Skill Set Score	Grades				
	Overall	K	1	2	3
Phonemic Awareness Sub-Domain					
Rhyming and Word Families	0.84	0.75	0.82	0.84	0.86
Blending Word Parts	0.84	0.77	0.82	0.85	0.87
Blending Phonemes	0.84	0.76	0.82	0.85	0.86
Initial and Final Phonemes	0.85	0.74	0.81	0.83	0.83
Consonant Blends (PA)	0.84	0.75	0.82	0.84	0.86
Medial Phoneme Discrimination	0.85	0.73	0.80	0.81	0.80
Phoneme Isolation/Manipulation	0.85	0.74	0.81	0.83	0.84
Paragraph-Level Comprehension Sub-Domain					
Comprehension of Paragraphs	0.85	0.74	0.81	0.83	0.82
Phonics Sub-Domain					
Long Vowel Sounds	0.85	0.74	0.81	0.82	0.82
Short Vowel Sounds	0.85	0.75	0.81	0.83	0.83
Initial Consonant Sounds	0.84	0.75	0.82	0.85	0.86
Final Consonant Sounds	0.84	0.75	0.82	0.84	0.85
Sound-Symbol Correspondence: Consonants	0.84	0.76	0.82	0.84	0.86
Sound-Symbol Correspondence: Vowels	0.85	0.74	0.81	0.83	0.84
Word Families/Rhyming	0.85	0.74	0.81	0.83	0.84
Consonant Blends (PH)	0.85	0.75	0.82	0.84	0.84
Consonant Digraphs	0.85	0.75	0.81	0.83	0.84
Phonics Sub-Domain (continued)					
Variant Vowel Sounds	0.85	0.75	0.82	0.84	0.85
Other Vowel Sounds	0.85	0.74	0.81	0.83	0.84
Structural Analysis Sub-Domain					
Word Building	0.85	0.74	0.81	0.83	0.83
Words with Affixes	0.84	0.75	0.82	0.84	0.86
Syllabification	0.84	0.76	0.82	0.85	0.86
Compound Words	0.85	0.74	0.81	0.83	0.82
Sentence-Level Comprehension Sub-Domain					
Comprehension at the Sentence Level	0.85	0.74	0.81	0.83	0.84

Table 16: Skill Set Score Split-Half Reliability, Overall and by Grade (Continued)

Skill Set Score	Grades				
	Overall	**K**	**1**	**2**	**3**
Vocabulary Sub-Domain					
Word Facility	0.85	0.76	0.82	0.84	0.86
Synonyms	0.85	0.74	0.81	0.83	0.82
Antonyms	0.85	0.74	0.81	0.83	0.83
Visual Discrimination Sub-Domain					
Letters	0.84	0.79	0.84	0.85	0.87
Identification and Word Matching	0.84	0.76	0.82	0.85	0.87

Scaled Score SEMs

Three different sources of data were available for estimating the aggregate standard error of measurement of STAR Early Literacy Enterprise Scaled Scores:

1. The averages and standard deviations of the conditional SEM (CSEM) values calculated by the STAR Early Literacy Enterprise 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.
3. Estimates of the standard error of measurement based on differences between the initial test and retest Scaled Scores for those students who took the test twice.

For the test-retest score data, the standard deviation of the test score differences, divided by the square root of 2 was used to estimate the standard error of measurement.[5]

Table 17 presents three different sets of estimates of the STAR Early Literacy Enterprise measurement error: average conditional standard errors of measurement (CSEM), global standard error of measurement, and retest standard errors of measurement. Two points should be noted here:

1. SEMs calculated using the conditional SEM method probably understate measurement error, since they are based on IRT models that do not fit the

5. Assuming that (1) measurement error variance is the same for both the initial test and the retest, and (2) the measurement errors are uncorrelated, the variance of the score differences is two times the measurement error variance of either test. The standard error of measurement is therefore the square root of the variance of the score differences divided by 2, which is identical to the standard deviation of the difference divided by the square root of 2.

response data perfectly and assume that the IRT item difficulty parameters are estimated without error.

2. SEMS calculated using the test-retest score differences probably overstate the measurement error of a single STAR Early Literacy Enterprise administration, since these estimates are subject to the influence of individual variation over time, item sampling differences between the initial test and retest, and other factors not present in a single administration of the test.

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 17 provides estimates of the three SEM calculations for each grade and overall for all grades, Pre-K to Grade 3.

Table 17: STAR Early Literacy Enterprise Standard Errors of Measurement from 2014 Norming Sample

Grade	N	Conditional Standard Error of Measurement		Global Standard Error of Measurement	Retest Standard Error of Measurement
		Average CSEM	Standard Deviation		
Pre-K	48,078	44	13	45	63
K	48,078	49	9	50	77
1	48,078	44	11	45	73
2	48,078	37	14	39	62
3	48,078	35	14	37	60
All Grades	240,390	42	13	43	68

Validity

Test validity is often described as the degree to which a test measures what it is intended to measure. 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. 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 15, 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. Almost 2,400 of those items have been retained for use in STAR Early Literacy. In every administration of STAR Early Literacy, items measuring each of the seven literacy domains are used.

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 Enterprise scores to the kinds of external variables cited above.

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 has been obtained in both the Calibration Study and the Validation Study.

Calibration Study Data

Table 18 lists summary statistics for age and STAR Early Literacy Scaled Scores by school grade in the Calibration Study.

Table 18: Median Age and Scaled Score by Grade in the Calibration Study

		Median Values		
Grade	**N**	**Age**	**Scaled Score**	**Standard Deviation**
Pre-K	2,584	4.6	509	87
K	5,938	5.6	580	85
Grade 1	10,768	6.7	703	83
Grade 2	6,852	7.7	779	82
Grade 3	6,115	8.7	826	63

As these data indicate, scores from the STAR Early Literacy Calibration Study do show the expected pattern of relationship to age and grade level—scores increase systematically from pre-kindergarten through grade 3. The standard deviation statistics show that score variability was similar from pre-kindergarten through grade 2, but there was much less variability in grade 3.

Validation Study Data

Table 19 lists summary statistics for age and STAR Early Literacy Scaled Scores by school grade in the Validation Study.

Table 19: Median Age and Scaled Score by Grade in the Validation Study

		Median Values		
Grade	**N**	**Age**	**Scaled Score**	**Standard Deviation**
Pre-K	610	5.1	426	104
K	2,323	6.1	576	108
Grade 1	2,578	7.2	749	115
Grade 2	1,945	8.2	813	90
Grade 3	2,063	9.2	838	76

As was true of the non-adaptive Calibration Study data, adaptive test scores from the Validation Study increased systematically from pre-kindergarten through grade 3. The standard deviation statistics show that score variability was similar from pre-kindergarten through grade 1, but less variable in grades 2 and 3.

The Validation Study took place during April and May, the seventh and eighth months of the school year. The data in Table 19 therefore represent ages and Scaled Scores about three-quarters of the way through the school year.

Figure 4 illustrates the relationship of Scaled Scores to school grade in a more complete way. The figure consists of curves that display the percentile equivalents of Scaled Scores separately for each grade from pre-kindergarten through grade 3. The point of showing these figures together is to emphasize that 1) score distributions in these five grades are quite different from one another; and 2) grade-to-grade differences are largest at the lowest grades, and become considerably smaller by grade 3, by which time most students can be expected to have mastered early literacy skills.

Figure 4: STAR Early Literacy Scaled Score Percentiles by Grade as Observed in the US Sample in the Validation Study

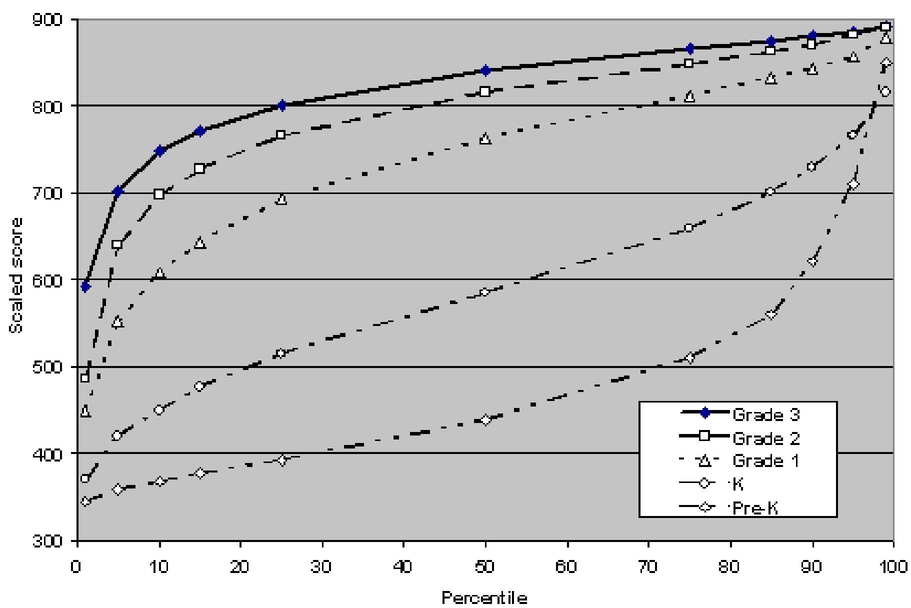
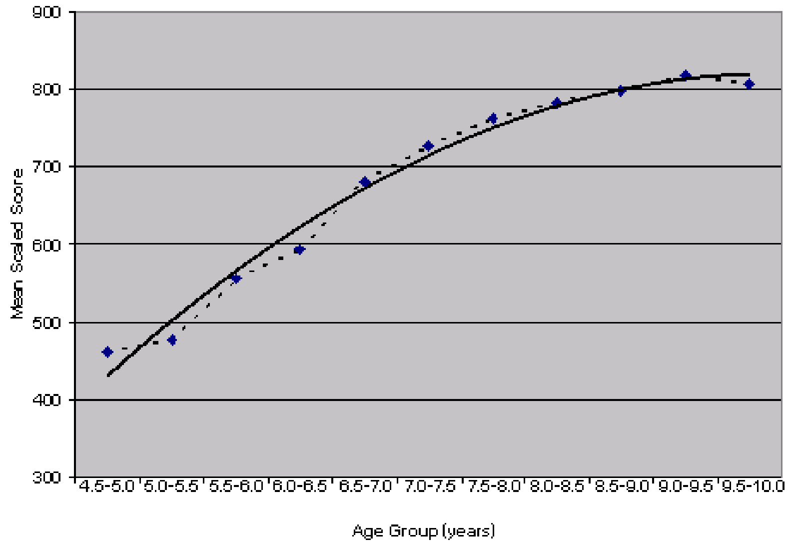


Figure 5 displays average STAR Early Literacy Scaled Scores for 11 age groups of students participating in the Validation Study. Each age group spans a six-month age range; the youngest group ranged from 4.5 to 5.0 years old on the date of the STAR Early Literacy assessment; the oldest group ranged from 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 5: STAR Early Literacy Scaled Scores as a Function of Age; Mean Scaled Scores for 11 Age Groups in the Validation Study



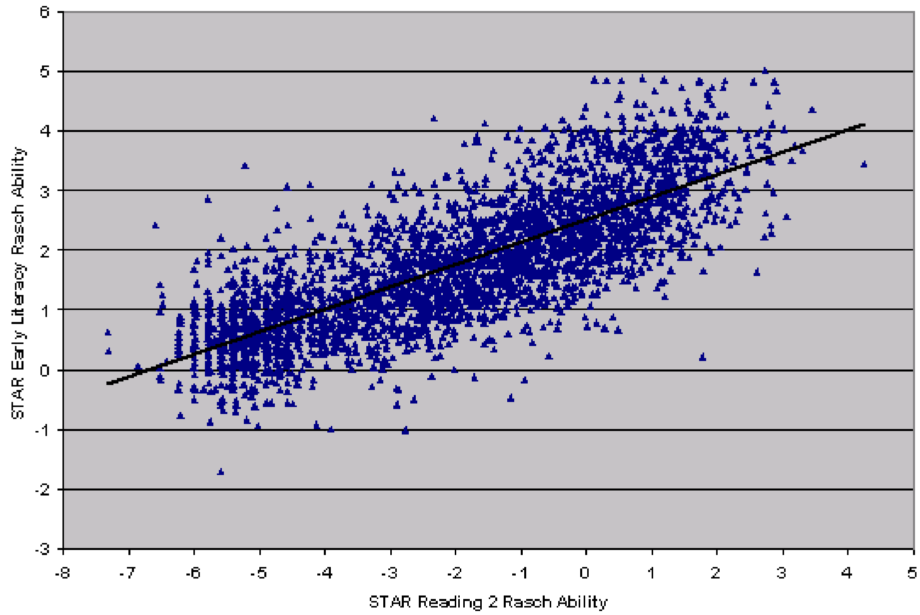
Relationship of STAR Early Literacy Scores to Other Tests

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 highly with measures of reading, literacy, and readiness. To evaluate this, standardized reading and other test scores were collected for some of the students participating in the Calibration Study and in the Validation Study. In the Calibration Study, STAR Reading 2.1, a computer-adaptive reading test, was administered specifically for this purpose. In the Validation 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 publication of STAR Early Literacy, additional data have been collected that show the relationship of STAR Early Literacy scores to scores on other tests. Below, results from the Calibration Study are presented first, followed by results from the Validation Study, and then from recent studies.

Calibration Study Results

During the Calibration Study, over 3,000 students in grades 1 through 3 took STAR Reading 2.1 in addition to the STAR Early Literacy tests. Figure 6 shows a plot of STAR Early Literacy Enterprise Rasch ability scores against STAR Reading 2.1 Rasch ability scores.

Figure 6: 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 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. 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.

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 (see Table 20). 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 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.

Tables 20 and 21 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 20 displays "concurrent validity" data, that is, correlations observed when two test scores and other tests were administered at close to the same time. Table 21 provides validity estimates with external tests given prior to STAR Early Literacy administration in spring 2001. Table 22 provides the predictive validity estimates with criterion tests given well after STAR Early Literacy.

Tables 20, 21, and 22 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 20: 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*	–	–	–	–	–	–

Table 20: Concurrent Validity: STAR Early Literacy Correlations with Tests Administered in Spring 2001, Grades K–3[a] (Continued)

Test Form	Date	Score	K		1		2		3	
			n[b]	r	n	r	n	r	n	r
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*
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.”

Table 21: 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	–	–	–	–	–	–
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*

Table 21: Other External Validity Data: STAR Early Literacy Correlations with Tests Administered Prior to Spring 2001, Grades K–3[a] (Continued)

Test Form	Date	Score	K		1		2		3	
			n[b]	r	n	r	n	r	n	r
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 22: 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 06[P]	–	–	1,945	0.47*	685	0.64*	30	0.90*	–	–
Fall 05	Spr 07[P]	–	–	1,945	0.42*	685	0.62*	30	0.72*	–	–
Spr 06	Fall 06[P]	22	0.67*	1,945	0.58*	685	0.77*	30	0.85*	–	–
Spr 06	Spr 07[P]	22	0.50*	1,945	0.59*	685	0.71*	30	0.71*	–	–

Table 22: Predictive Validity Data: STAR Early Literacy Predicting Later Performance for Grades Pre-K–3 (Continued)

Predictor Date	Criterion Date[a]	Pre-K[b]		K		1		2		3	
		n	r	n	r	n	r	n	r	n	r
STAR Reading											
Fall 03	Fall 05[P]	–	–	671	0.49*	698	0.58*	194	0.65*	–	–
Win 04	Fall 05[P]	–	–	671	0.54*	698	0.62*	194	0.61*	–	–
Spr 04	Fall 05[P]	–	–	671	0.73*	698	0.67*	194	0.65*	–	–
Fall 03	Win 06[P]	–	–	552	0.43*	653	0.56*	469	0.64*	–	–
Win 04	Win 06[P]	–	–	858	0.55*	772	0.61*	227	0.57*	–	–
Spr 04	Win 06[P]	–	–	639	0.51*	551	0.66*	254	0.59*	–	–
Fall 03	Spr 06[P]	–	–	282	0.47*	376	0.61*	291	0.62*	–	–
Win 04	Spr 06[P]	–	–	497	0.56*	428	0.59*	167	0.59*	–	–
Spr 04	Spr 06[P]	–	–	480	0.55*	343	0.58*	195	0.57*	–	–

Summary

Grades	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 ($p < 0.05$).

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 63 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 23. 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.

Table 23: 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

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 63 correlations reported in Tables 20 and 21, if the true validity were zero, is virtually zero. Because the 63 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.

Post-Publication Study Data

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 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.

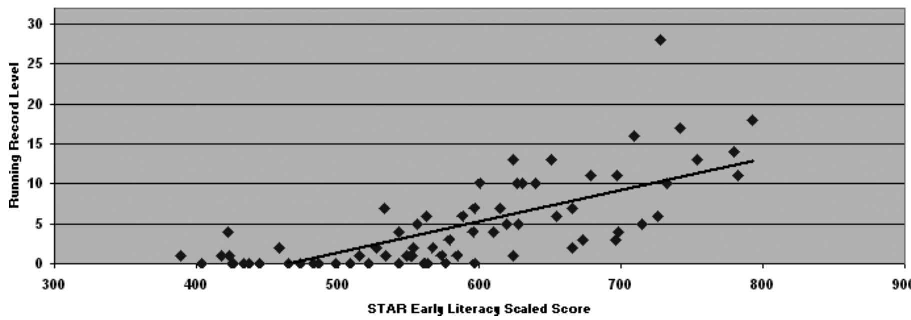
Running Record

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. Figure 7 shows a scatterplot of Running Record scores ranging from 0 to 29 against STAR Early Literacy scale scores ranging from 389 to 793. 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.

Figure 7: Running Record vs. STAR Early Literacy Scaled Scores Kindergarten and Grade 1 Data, January 2002; Correlation Coefficient = 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. MLPP tests are generally administered one-on-one by a teacher or other trained test administrator. 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. MLPP tests are not normed. The MLPP system consists of several tests. Each school may select the tests it wishes to use. In addition, it may substitute other tests that are not traditionally a part of the MLPP. Two Michigan elementary schools that use both STAR Early Literacy and the Michigan Literacy Progress Profile participated in a study of the relationship between the two assessments.

Two Michigan elementary schools that use both MLPP and STAR Early Literacy provided data for the 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 revealed statistically significant and generally high correlations at significant levels in both Fall 2003 and Spring 2004.

As shown in Table 24, 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; all are statistically significant. 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[6] on the MLPP tests. Ceiling effects affected all but two of the Spring 2004 MLPP tests (Word Lists and Known Words.)

To estimate what the correlation would have been had there been no ceiling effect on the MLPP tests, the McNemar correction formula was applied. It takes into account the variance of scores in both Fall and Spring. The correction was not applied to correlations with the two MLPP tests for which the ceiling effect was not evident.

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. To request a free copy, call Renaissance Learning at 1-800-338-4204.

Table 24: STAR Early Literacy Scaled Score Correlations with Michigan Literacy Progress Profile Raw Scores, Combined K and 1

Test	Fall 2003		Spring 2004	
	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

6. Ceiling effects occur when the overall ability level of students exceeds the difficulty level of the test items, resulting in a score distribution with a large proportion of the students attaining the highest possible scores on the test. Ceiling effects attenuate the magnitudes of correlation coefficients by reducing test score variance. A statistical "correction for range restriction" is intended to correct this.

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,[7] TPRI,[8] and GRADE.[9] 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.

DIBELS—The following tests were administered at the grades indicated: Initial sound fluency (ISF), letter naming fluency (LNF), phoneme segmentation fluency (PSF), nonsense word fluency (NWF) and word usage fluency (WUF) were administered to kindergartners. First graders took letter naming fluency (LNF), phoneme segmentation fluency (PSF), nonsense word fluency (NWF), word usage fluency (WUF), oral reading fluency (ORF), and retell fluency (RF). Second graders took just four assessments: nonsense word fluency (NWF), word usage fluency (WUF), oral reading fluency (ORF), and retell fluency (RF). At their discretion, some teachers omitted specific assessments.

GRADE—Kindergarten students took Word Reading and 5 subtests measuring phonological awareness and some phonics skills. All students in grades 1 and 2 took Word Reading, Word Meaning, Sentence Comprehension, and Passage Comprehension. GRADE reports subtest raw scores, and composite standard scores and scale scores.

TPRI—All students took one, two, or three short screening subtests: three screening tests for kindergarten, two for grade 1, and one for grade 2. The screener was followed by short inventory tests (tasks) measuring specific phonemic awareness, graphophonemic knowledge, and fluency and/or comprehension skills. In TPRI, the choice of inventory tests is made adaptively, using branching rules followed by the teacher; not every student took all subtests, but all should have taken the comprehension and/or fluency measures. A listening comprehension test is used at kindergarten. At grades 1 and 2, students read a leveled oral reading fluency passage, followed by a short reading comprehension test on the passage; the choice of the reading passage is based on the student’s performance on a 15-item word reading task.

---

7. Dynamic Indicators of Basic Early Literacy Skills (University of Oregon, Institute for Development of Educational Achievement, 2002).
8. Texas Primary Reading Inventory 2004–2006 (Texas Education Agency and the University of Texas System, 2003).
9. Group Reading Assessment and Diagnostic Evaluation (American Guidance Service, Inc., 2001).

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 25.[10]

As the data in Table 25 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.

Table 25: 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

10. 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 25: STAR Early Literacy Correlations with DIBELS, GRADE, and TPRI (Continued)

	Kindergarten		Grade 1		Grade 2	
	r	n	r	n	r	n
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
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	–	–

Table 25: STAR Early Literacy Correlations with DIBELS, GRADE, and TPRI (Continued)

	Kindergarten		Grade 1		Grade 2	
	r	n	r	n	r	n
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

Predictive Validity

An internal study by 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.

It also provided a longer-term analysis of the level of predictive validity of early reading skills relative to later reading skills from all three time points: from kindergarten to second grade, first grade to third grade, and second grade to fourth grade.

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. The six cohorts of students were those that started kindergarten, first grade, or second grade in 2001–2002 and in 2002–2003.

Demographic data are not available on the students themselves. However, data are available on the demographic makeup of the schools they attended. The average school size was about 490 with a standard deviation of about 210 students. The ethnic distribution of the sample was about 59% European American, 14% African American, 22% Hispanic American, 3% Native American, and 2% Asian American.

To minimize the cohort effect on any one estimate, the two comparable cohorts from successive school years were combined. For example, the students starting kindergarten in the 2001–2002 school year were combined with those starting kindergarten in the 2002 school year. This first cohort had second grade reading scores two years later; that is, for the 2003–2004 school year. The second cohort had second grade reading scores for the 2004–2005 school year. Thus, the six cohorts were combined into three age-similar groups defined by their initial grade level: a K-grade 2 group, a grade 1-grade 3 group, and a grade 2-grade 4 group.

Each cohort's early literacy skills were assessed during the fall, winter, and spring of the beginning school year using STAR Early Literacy. The fall assessment period took place in August–September, the winter assessment in December–January, and the spring assessment in April–May. The final criterion variable was the score on the STAR Reading test taken during the fall assessment period two years after the initial early literacy assessment. Two complete school years separated the first predictor variable measurement and this criterion measurement.

Since the students were assessed using STAR Early Literacy at three time points within their initial grade levels, and if that test has developmental validity, then scores should increase at each successive time point and the correlations across occasions should be substantial. In addition to each cohort's scores increasing with time, STAR Early Literacy scores should also increase from grade to grade across cohorts.

The use of the aggregated results will provide a measure of the general validity of the scores in predicting later reading scores. This permits the assessment of predictive validity across multiple time frames. Breaking the total sample down into three groups by initial grade level status, one can analyze the predictive validity of previous years' fall, winter, and spring STAR Early Literacy scores relative to STAR Reading scores at the second grade.

Means and standard deviations of scale scores at each measurement time point are provided in Table 26. For STAR Early Literacy tests in the initial year, overall average scale scores across the school year increased, indicating that the early literacy assessment appears to follow an assumed developmental trajectory. This is also evident for scores within each initial grade level: kindergarten and first and second grades.

Table 26: Descriptive Statistics for Each Measurement Occasion for the Total Sample and for Each Grade-Level Group

Grade Groups		STAR Early Literacy Initial-Year Scale Scores Fall	Winter	Spring	STAR Reading End-Year Scale Scores[a] 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
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. STAR Reading was taken two years after the initial-year fall administration of STAR Early Literacy.
b. Abbreviations: STD Dev = standard deviation; N = number of students; ρxx = reliability estimate.
c. The group with initial grade in kindergarten.
d. The group with initial grade in 1st grade.
e. The group with initial grade in 2nd grade.

To analyze whether differences between grade levels were statistically significant, a MANOVA was used with the three STAR Early Literacy scores as outcomes and grade level fixed. Results indicated significant differences existed, Wilks' Lambda = 0.5465, $F(6, 3404) = 199.19$, $p < 0.001$. Follow up analysis indicated significant differences existed between all grades at all measurement occasions, fall STAR Early Literacy scores, $F(2, 2727) = 1302.84$, $p < 0.001$, $\eta^2 = 0.49$, winter scores, $F(2, 2975) = 1005.79$, $p < 0.001$, $\eta^2 = 0.40$, and spring scores, $F(2, 3381) = 1083.23$, $p < 0.001$, $\eta^2 = 0.39$. All differences between grade levels were also significant, which indicated that at each measurement occasion, the higher grade scored significantly higher than the lower grade.

An interesting phenomenon was noted when evaluating the differences between spring scores of one school year on STAR Early Literacy and the next higher grade level's fall scores on STAR Early Literacy: fall scores for the next higher grade were slightly lower than spring scores for the preceding grade. For instance, the average scale score of about 615 at the end of kindergarten (spring) is about one point higher than the average of first grade students' scores at the beginning of the school year (fall). A larger difference is seen between the average spring first grade scale score of about 745 and the fall second grade average of about 725. This does not necessarily invalidate claims of developmental validity, because the drop in scores coincides with the occurrence of summer break. This summer break from school has been identified with regression in student scores between grades (Allington & McGill-Franzen, 2003; Bracey, 2002; Malach & Rutter, 2003; McGill-Franzen & Allington, 2003). The drop in scores between grade levels is consistent with the research just cited; that is, it does not necessarily represent an inversion or discontinuity in measurement, but rather an empirical phenomenon sometimes referred to as the summer slump.

Table 27 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. These results are consistent for each of the subgroups partitioned by initial grade level.

Table 27: Validity Coefficients[a] 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 Level Group	Administration Time		STAR Early Literacy[b]		STAR Reading[c]	
			Winter	Spring	Fall	Corrected[d]
Total Sample	Fall	r[e]	0.75	0.69	0.64	0.71
		N[f]	1,265	1,382	1,740	
	Winter	r	–	0.78	0.67	0.76
		N	–	2,022	2,182	
	Spring	r	–	–	0.70	0.79
		N	–	–	2,810	
K to 2nd	Fall	r	0.52	0.50	0.47	0.56
		N	511	543	572	
	Winter	r	–	0.64	0.57	0.70
		N	–	922	880	
	Spring	r	–	–	0.59	0.73
		N	–	–	1,095	
1st to 3rd	Fall	r	0.66	0.56	0.54	0.67
		N	600	663	811	
	Winter	r	–	0.68	0.62	0.76
		N	–	881	1,012	
	Spring	r	–	–	0.66	0.78
		N	–	–	1,255	
2nd to 4th	Fall	r	0.56	0.58	0.51	0.61
		N	154	176	357	
	Winter	r	–	0.75	0.55	0.64
		N	–	219	290	
	Spring	r	–	–	0.58	0.68
		N	–	–	460	

a. All coefficients are statistically significant ($p < 0.001$).
b. STAR Early Literacy Enterprise 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.
f. "N" indicates the number of students used to calculate the validity coefficient.

In summary, the data displayed in Tables 26 and 27 of this section 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.

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) for interim assessments were contacted and asked to participate 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. In addition, schools were asked to submit fall, winter, and spring interim assessment data from the previous school year, and any student that had a valid STAR Early Literacy and DORF assessment within a 2-week time span was used in the analysis. Thus, the research involved both a current sample of students and also historical data from those same schools. No schools assessed 1st-grade students in the fall on the DIBELS passages, so there was no fall data for grade 1 in the analysis.

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 a 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. Within each grade, correlations between the DORF WCPM score and the underlying STAR Early Literacy Rasch score for each student were calculated to get an estimate of the relation between the two measures.

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. The sample was composed of 61% of students of European ancestry; 21% of African ancestry; 11% of Hispanic ancestry; and the remaining 7% of Native American, Asian, or other ancestry. About 3% of the students were eligible for services due to limited English proficiency (LEP), and about 14% were eligible for special education services.

Descriptive statistics and correlations between the STAR Early Literacy Rasch scores and DORF raw score (measured in words read correctly within a minute [WCPM]) are provided in Table 28. All correlations were statistically significant ($p < 0.01$). Scatterplots of the relationship between STAR Early Literacy Rasch scores and WCPM on the benchmark passages of DIBELS are shown in Figures 8–10.

Table 28: Correlations between STAR Early Literacy and DIBELS Oral Reading Fluency

Grade	N	STAR Early Literacy Rasch Score		DIBELS WCPM		Correlation
		Mean	SD	Mean	SD	
1	2,028	1.38	0.96	37.96	27.70	0.68
2	729	1.53	0.93	49.20	27.43	0.63
3	464	2.18	1.06	71.06	31.93	0.65

Figure 8: Scatterplot of Grade 1 DIBELS Oral Reading Fluency Score (DORF WCPM) and STAR Early Literacy Rasch Score

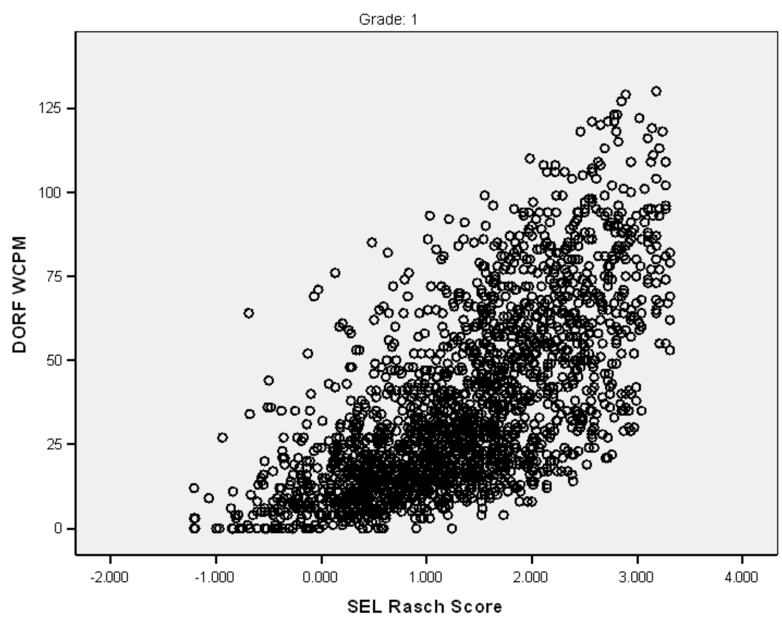


Figure 9: Scatterplot of Grade 2 DIBELS Oral Reading Fluency Score (DORF WCPM) and STAR Early Literacy Rasch Score

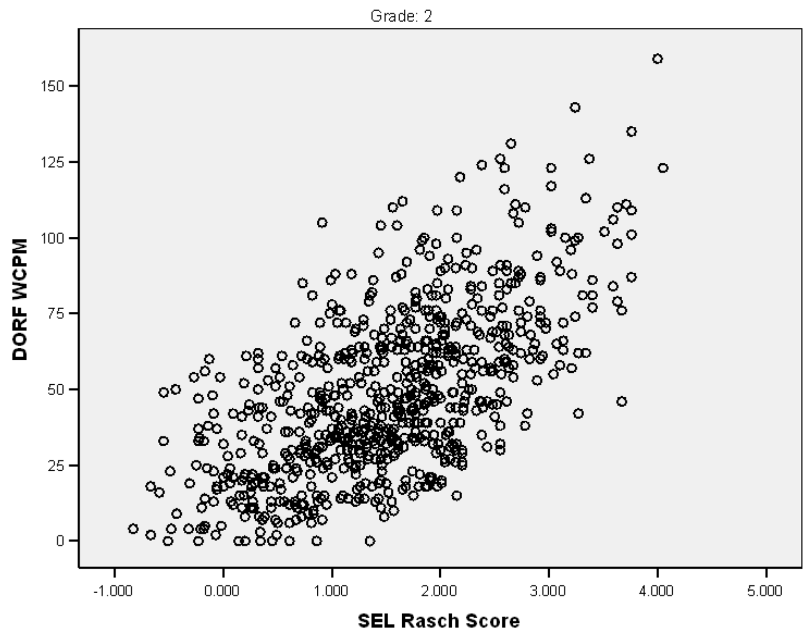
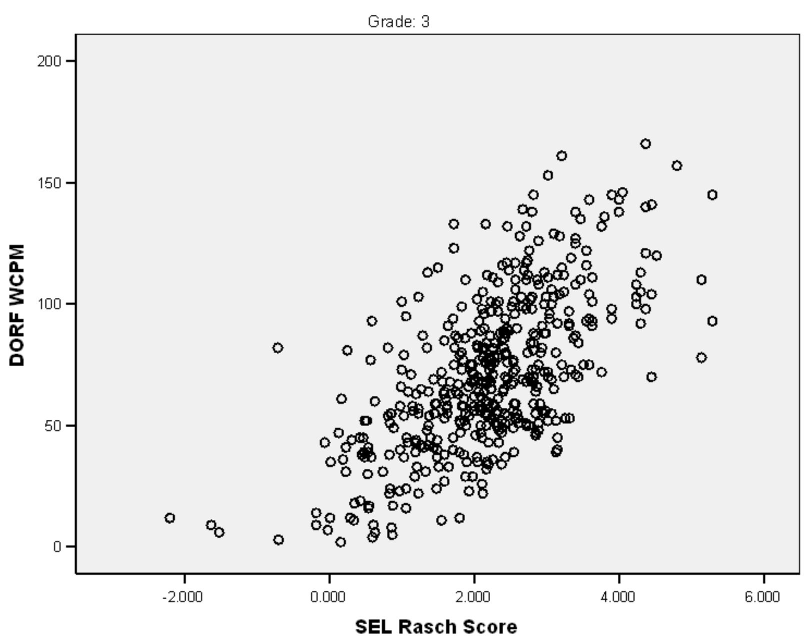


Figure 10: Scatterplot of Grade 3 DIBELS Oral Reading Fluency Score (DORF WCPM) and STAR Early Literacy Rasch Score



Correlations between the Est. ORF and DORF WCPM are displayed in Table 29 along with the mean difference, standard deviation of differences, and the 95% CI of the mean difference between the estimated score (Est. ORF) and the observed score (WCPM). 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 29: Relations STAR Early Literacy Est. ORF and DIBELS Oral Reading Fluency

Grade	Correlation	Mean Difference (95% CI)	SD Difference	t-test
1	0.69	3.99 (3.07, 4.90)	20.99	$t(2027) = 8.56$, $p < 0.001$
2	0.64	–3.67 (–5.37, –1.97)	23.40	$t(729) = -4.23$, $p < 0.001$
3	0.64	–7.29 (–9.82, –4.76)	27.77	$t(463) = -5.67$, $p < 0.001$

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 Enterprise version. However, because of the similarity of the pre-Enterprise and Enterprise versions, and the high degree of correlation between them reported here, there is ample reason to consider the Enterprise version of STAR Early Literacy to be equivalent to other versions, and therefore to accept evidence of the validity of those earlier versions as being applicable to the Enterprise version as well. Indeed, the observed correlation between the Enterprise and earlier "service" version, corrected for attenuation due to measurement error, is nearly perfect. Along with the close correspondence of Enterprise-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 STAR Early Literacy Enterprise, and can accept all of the summary statements here as equally applicable to the Enterprise 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 presented in Table 20 on page 61 and Table 21 on page 63. These 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. The Meta-Analysis section 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.

Validation Research Study Procedures

The Validation Research Study

The technical results of the STAR Early Literacy Calibration Study were excellent, with the tests showing good measurement properties, a high degree of reliability, and high correlation with an independent measure of reading ability. However, the Calibration Study was conducted using conventional tests, while STAR Early Literacy was designed to be an adaptive test.

Because the technical properties of the adaptive version may be somewhat different from those found in the Calibration Study, additional psychometric research data were collected in the Spring of 2001 with the first computer-adaptive version of STAR Early Literacy. Data from this Validation Research Study were intended to assess a number of technical characteristics of the adaptive version, including the following:

- Reliability and measurement precision of the adaptive STAR Early Literacy tests.
- Score distributions by age and grade.
- Validity of STAR Early Literacy.
- Appropriateness of the adaptive version of STAR Early Literacy.
- Teacher reactions to the design of the assessment.

Sample Characteristics

The Validation Study took place in the Spring of 2001. Although the Validation Study sample was targeted to include schools using certain standardized early literacy and reading tests, the participating school districts, specific schools, and individual students were approximately representative of the US school population, in terms of the following three key variables:

- **Geographic Region:** Using the categories established by the National Education Association, schools fell into four regions: Northeast, Midwest, Southeast, and West.
- **School System and Per-Grade District Enrollment:** Statistics distributed by MDR (2001) identified public and nonpublic schools. Public schools were categorized into four groups based on their per-grade district enrollment: fewer than 200 students, 200–499 students, 500–1,999 students, and more than 1,999 students.
- **Socioeconomic Status:** Using the Orshansky Indicator from MDR (2001), public schools were categorized based on the proportion of students in the district who fall below the federal poverty level. As a result, schools were identified as being either of High, Average, or Low socioeconomic status. (Nonpublic schools were not classified by socioeconomic status, as socioeconomic data were not available for them.)

These factors provide a sampling frame comprising a 52-cell matrix (4 regional zones × 4 public school enrollment groups × 3 socioeconomic categories, plus 4 regional cells for nonpublic schools). All schools in the US were categorized into one of the 52 cells, and participation was requested from sufficient numbers of schools to complete the desired sample. This US sampling frame was used only for reference in the analysis of the Validation Study results; the sample itself was recruited principally from among schools that had participated in the Calibration Study described in "Core Progress Learning Progression for Reading and the Common Core State Standards" on page 36. In addition to US schools, schools in Canada were also recruited to participate in the Validation Study.

In April 2001, the 101 schools that agreed to participate received a version of STAR Early Literacy designed to gather the validation research data. This version of the program captured the test scores and item responses for each of the students participating.

The final Validation Study sample included approximately 11,000 students from 84 schools in the US and Canada (Appendix A lists the name, location, and region of every school that participated in this and other research phases).

Participating schools were asked to administer the tests within a 4-week window spanning April and May of 2001. In order to provide test-retest reliability data, schools were asked to test every student twice, with an interval of one to seven days between sessions. In all, 10,624 students in grades from pre-kindergarten through grade 3 took the adaptive version of STAR Early Literacy, and 9,236 of them took it two or more times.

Table 30 compares US student sample characteristics against percentages in the US population.

Table 30: Sample Characteristics, STAR Early Literacy Validation Study, Spring 2001 (N = 9,038 US Students)

		Students	
		National %	**Sample %**
Geographic Region	Northeast	20.4	0.2
	Midwest	23.5	26.1
	Southeast	24.3	44.5
	West	31.8	29.1
District Socioeconomic Status	Low	28.4	32.5
	Average	29.6	54.2
	High	31.8	11.6
	Not Available[a]	10.2	1.7
School Type and District Enrollment	Public		
	< 200	15.8	34.8
	200–499	19.1	25.9
	500–1,999	30.2	23.7
	> 1,999	24.7	13.8
	Non-Public	10.2	1.7

a. Socioeconomic Status data were not available from non-public schools.

In addition to the sample characteristics summarized in Table 30, additional information about participating schools and students was collected. This information is summarized in Table 31, Table 32, and Table 33. These tables also include national figures based on 2001 data provided by MDR.

Table 31: School Locations, STAR Early Literacy Validation Study, Spring 2001 (N = 71 US Schools, 9,038 US Students)

	Schools		Students	
	National %	**Sample %**	**National %**	**Sample %**
Urban	27.8	19.7	30.9	16.9
Suburban	38.3	29.6	43.5	36.8
Rural	33.2	50.7	24.8	46.2
Unclassified	0.7	0.0	0.7	0.0

Table 32: Non-Public School Affiliations, STAR Early Literacy Validation Study, Spring 2001 (N = 2 US Schools, 157 US Students)

	Schools		Students	
	National %	**Sample %**	**National %**	**Sample %**
Catholic	39.7	50.0	51.8	86.0
Other	60.3	50.0	48.2	14.0

Table 33: Ethnic Group Participation, STAR Early Literacy Validation Study, Spring 2001 (N = 9,038 US Students)

	Students	
Ethnic Group	**National %**	**Sample %**
Asian	3.4	0.5
Black	14.5	12.8
Hispanic	12.7	6.0
Native American	0.9	0.1
White	54.7	38.0
Unclassified	13.8	42.8

Test Administration

The adaptive tests drew their test items from a bank of 2,435 items[11] chosen following the Calibration Study. Each student's test consisted of 25 items, selected one at a time contingent on the student's ability estimate, which was updated after each item. The selection of the initial test item was based on an initial ability estimate that varied as a function of grade placement and the student's age at the time of the test. For retests, the initial ability estimates were determined as they were for the student's initial test; however, items administered during the student's initial test were not administered again during retests.

Each 25-item test consisted of two parts, arranged with the intention of controlling test duration: Items 1 through 15 were drawn from a subset of items known to have relatively short administration times; items 16 through 25 were drawn from among items with longer administration times.[12] Content balancing specifications governed each of the two parts, ensuring that every test included a specific number of items from each of the seven literacy domains. No items requiring the student to read sentences or paragraphs were administered at the lowest two grades (pre-kindergarten and kindergarten).

Data Analysis

After the participating schools tested their students, they returned their student test data on floppy disks for analysis. In all, there were 10,624 students with complete, first administration STAR Early Literacy tests. Prior to data analysis, all test records were screened for quality control purposes. In a few cases, discrepancies were found between the student's grade and the initial difficulty of the test. In the interests of maximal standardization, such cases were omitted from the analyses. A few other cases were identified in which the detailed test records strongly suggested unmotivated performance; these cases were likewise excluded from some or all analyses. After completion of all data screening, 10,061 first administration cases and 9,146 retest cases proceeded to the analyses.

11. Subsequent to the Validation Study, the size of the adaptive item bank was further reduced to 2,350 items used for testing, with 18 more items reserved for use as practice items. The item count in the bank stood at 2,351 until it was reduced to 2,350 with the release of STAR Early Literacy RP version 2.3.

12. In version 2.x Renaissance Place and higher of STAR Early Literacy, 16 items are administered in the first part of the test, and 9 items are administered in the second part.

Table 34 presents the STAR Early Literacy Scaled Scores summary data by grade for the US and Canadian samples separately.

Table 34: Summary of Scaled Score Statistics, STAR Early Literacy Validation Study

	Sample Size		Scaled Score Means		Scaled Score SDs		Scaled Score Medians	
Grade Level	**US**	**Canada**	**US**	**Canada**	**US**	**Canada**	**US**	**Canada**
Pre-K	449	78	472	469	112	75	439	467
K	1,982	325	588	570	103	107	585	564
1	2,423	400	742	677	94	112	763	688
2	1,769	371	796	782	79	77	816	800
3	1,905	359	823	815	63	71	841	834

More detailed score distribution data, including distributions of Domain and Skill Scores as well as Scaled Scores, are presented in “Score Definitions” on page 110. Other results from the Validation Study, including reliability and validity data, are presented in “Reliability and Measurement Precision” on page 44 and “Validity” on page 55.

STAR Early Literacy Enterprise Research Study Procedures

The STAR Early Literacy Enterprise version was completed early in 2012, by the incorporation of the Enterprise assessment blueprint and the Enterprise version of the adaptive item bank into application software to administer and score the Enterprise assessments. That marked the first point at which it was possible to administer computerized adaptive versions of the STAR Early Literacy Enterprise edition. What remained at that point was to collect evidence of the new assessment’s reliability, its psychometric equivalence to the previous version of STAR Early Literacy, and its validity as a measure of early literacy. The evidence to address those issues was collected in a specially designed research study early in 2012. The remainder of this section describes the research study itself, the data collected, and the results of analyses of those data.

The Research Study

The research study involved collecting three different data elements on a large sample of STAR Early Literacy students:

1. Scores on the new STAR Early Literacy Enterprise edition tests.
2. Scores on the previous version of STAR Early Literacy, referred to below as the "service" version.
3. Teachers' ratings of the students' mastery of a hierarchy of 10 early literacy skills, aligned to the Common Core State Standards (CCSS), and graduated in content and difficulty from pre-kindergarten through 3rd grade level.

Schools from throughout the US were recruited to participate in the research study. The intent was for each participating student to take both the service and Enterprise versions of STAR Early Literacy, on different days and in counterbalanced order of administration, and to be rated on the 10 CCSS skills by their teachers, independently of their performance on the tests themselves. All data on the Enterprise edition were to be collected during a one-month period spanning mid-February through mid-March 2012.

Sample Characteristics

Fifty schools from the US and Canada participated in the research study. In those schools, a total of 7,420 students took the Enterprise edition of the test; teachers completed the literacy skills ratings for 6,720 students. All students were also to take the service version; some students took the service version more than once during February and March 2012. The following data were included in files describing the schools:[13]

- Country
- Region (US only)
- State (US) or province (Canada)
- Grade range

Additionally, the following data were included in files describing the students:[14]

- Grade
- Age
- Gender
- Race/ethnicity

13. Observations on these data elements were missing for some schools.
14. Observations on some of these data elements were missing in many cases from student records.

Test Administration

Most of the participating schools were current users of the service edition of STAR Early Literacy; some were already using it in progress monitoring programs involving frequent administration of that test, as often as weekly in some cases. Participating schools were asked to administer the Enterprise and service versions of the test in counter-balanced order. Teachers were asked to complete the 10-item literacy skills ratings on their students early in the month-long test administration window.

Data Analysis

Data analysis focused on evaluating the following:

1. the equivalence of STAR Early Literacy Enterprise and service versions
2. other evidence of the validity of the Enterprise version

Equivalence and Validity of STAR Early Literacy Enterprise and Service Versions

The principal evidence related to the equivalence of the Enterprise and previous version consisted of score distributions and correlations. Three 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.
- **Percentile ranks.** Although the STAR Early Literacy norming study was completed in the summer of 2014, percentile ranks relative to the 2001 validity study sample were recorded and used for some purposes.

Validity was evaluated primarily by analyses of the statistical correlations between STAR Early Literacy Enterprise scores and the following external variables:

- Scores on the service version of STAR Early Literacy, which can be considered to be an alternate (but not parallel) test. Because some students took the service version on multiple occasions during the research window, their average scores were used in these analyses; averages of two or measurements are generally more reliable than single measures. There were records of 7,998 completed service version tests; mean service version scores were available for 7,152 students. After matching Enterprise scores to the same students' average service version scores, there were 7,070 matched pairs of scores available for correlational analysis.

- Students' age and grade. Age was available only for the 4,421 students whose dates of birth were recorded; grade was available for all 7,404 students with completed Enterprise tests. Data were available for students in grades pre-K through 5; however, the samples sizes were too small for analysis in any grades except Kindergarten through 3.
- Teachers' ratings of 6,720 students on the 10 CCSS-aligned literacy skills. Scores on the rating survey were for all of those students.

Results

Equivalence of STAR Early Literacy Enterprise and Service Versions

Table 35 lists summary statistics on Enterprise and service scores—including Rasch ability and Scaled Scores—for all grades combined, including grades pre-K through grade 5. The scores for the service 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 Enterprise Scaled Score was 755, 13 points higher than the average Scaled Score on the service version, 742. Similarly, the mean Enterprise Rasch ability score was 1.88, which was 0.18 logits (Rasch scale units) higher than the service version average of 1.70.

Table 35: Summary Statistics on Enterprise and Service Tests, All Grades Combined

Variable	N	Mean	Standard Deviation
Enterprise Scaled Score	7,070	755	104
Enterprise Rasch Score	7,070	1.88	1.37
Average Service Scale Score	7,070	742	106
Average Service Rasch Score	7,070	1.70	1.33

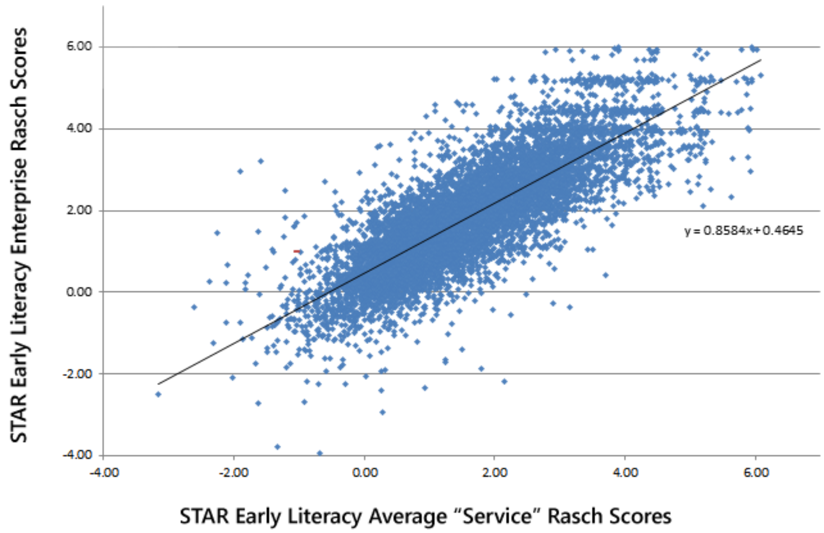
Table 36 displays the intercorrelations of the Enterprise test scores with the service version scores for all grades combined. The correlation between the Scaled Scores was 0.78; between the Rasch ability scores, the correlation was 0.80. All correlations were statistically significant ($p < 0.0001$).

Table 36: Correlations between Enterprise Test Scores and Average Service Version Scores, All Grades Combined, N = 7070

	Average Service Scaled Score	Average Service Rasch Score
Enterprise Scaled Score	0.78	0.76
Enterprise Rasch Score	0.78	0.80

Figure 11 displays the scatterplot of the service version Rasch ability scores versus the Enterprise scores. The plot illustrates the strong linear relationship between scores on the two tests.

Figure 11: Plot of STAR Early Literacy Average "Service" Rasch Scores versus STAR Early Literacy Enterprise Rasch Scores



The magnitudes of these 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 37.

Table 37: Correlations between Enterprise and Service Version Scaled Scores and Rasch Scores, Corrected for Attenuation

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

The corrected correlations shown in Table 37 are high, but substantially less than 1.00. This implies that the Enterprise and service versions are measuring highly related, but not identical, constructs.

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

Equivalent STAR Early Literacy Service Rasch Ability Score =
0.9746 × Enterprise Rasch Score – 0.1350

Other Correlational Evidence of STAR Early Literacy Enterprise Validity

Other currently available evidence of the validity of STAR Early Literacy Enterprise scores consists of their correlations with student age, grade, and literacy skills as rated by teachers. Evidence for each of these will be presented in this section.

Because STAR Early Literacy, including the Enterprise version, reports its scores on a single, vertical scale that is applicable regardless of student age and grade, we would expect scores to increase with students' ages and grade levels, and therefore would also expect at least a moderate degree of correlation of Enterprise scale scores with both age and grade. Evidence that such is the case is presented below in Tables 38 and 39.

Table 38 presents data on the relationship of Enterprise scaled scores to student age, for students less than 10 years old. Table 39 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 40 lists the correlations between scaled scores and age (0.44) and grade (0.51.) Both correlations show a moderate degree of association and are statistically significant ($p < 0.0001$).

Table 38: Summary Statistics on Enterprise 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 39: Summary Statistics on Enterprise 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

Table 40: Correlation Coefficients of STAR Early Literacy Enterprise Scaled Scores with Student Age and Grade

	Student Age	Grade
Enterprise Scaled Score	0.44	0.51
Number of observations	4,121	7,315

Table 41 lists summary statistics for age and STAR Early Literacy Scaled Scores by school grade in the STAR Early Literacy Enterprise equivalence study, for those students for whom age was recorded. As was true of the non-adaptive Calibration Study data, as well as the STAR Early Literacy Validation Study data, adaptive test scores from the Enterprise Equivalence Study increased systematically from kindergarten through grade 3. The standard deviation statistics show that score variability was similar for kindergarten and grade 1, but less variable in grades 2 and 3.

Table 41: Median Age and Scaled Score by Grade in the Enterprise Equivalence Study

Grade	N	Median Age	Median Scaled Score	Standard Deviation
0	1,223	6.00	698	103
1	1,564	7.02	768	103
2	816	8.01	837	70
3	403	9.04	849	67

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 Enterprise version (SELE) in 2012, the test blueprint did not prescribe specific numbers of numeracy items to be administered. The Enterprise version's item bank contains more numeracy items than previous versions, and the SELE 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 each SELE 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 SELE test, and the majority of those students also took the previous version of the SEL test within a few days before or following the

Enterprise version. The factor analysis was based on data from 6,785 students who participated in the study and also took the earlier version; student records with complete data on both tests were used for the factor analysis.

A SELE test comprises 27 items administered adaptively. Sequentially, the first 22 items are EL items and last 5 are EN items. For this analysis, SELE 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 of a SELE test. 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.[15] In addition, there were test scores on the Service version of STAR Early Literacy and skills ratings for all students in the study. The skills ratings were provided by teachers of the participants.

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.
- Service—scores on the previous SEL version, called the "Service" 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 42. 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, $r = .59$, 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

15. 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.

between the EL17 scores and the EN5 scores was .64 which was greater than that between the EL5 scores and the EN5 scores (*r* = .59).

Table 42: Score Correlations (N = 6,438)

	EL17	EL5	EN5	Service	Rating
EL17	1				
EL5	0.73*	1			
EN5	0.64*	0.59*	1		
Service	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.

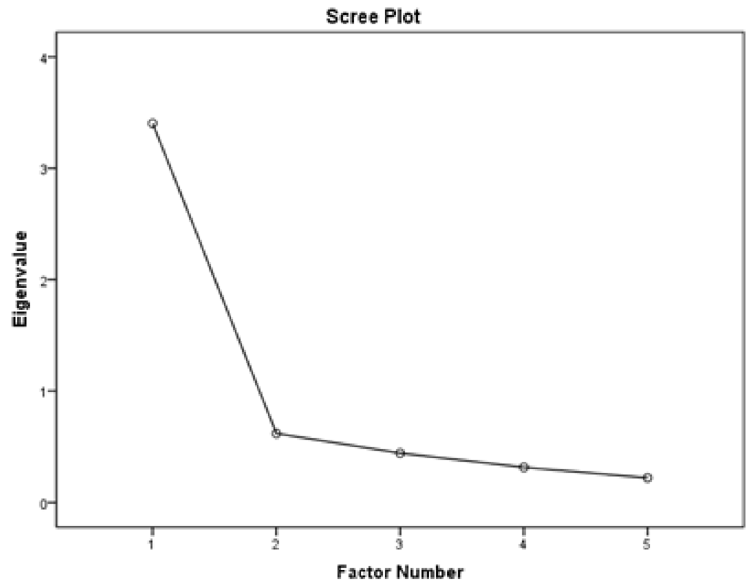
Often, correlations do not tell the whole story. Next, we present the results of the factor analysis that examined the relationship of the 5 variables to the underlying construct. Factor analysis can identify as many as 4 distinct latent attributes (factors) that account for the intercorrelations among 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 43. 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 43: Factor Loadings

Variable	Factor Loading
EL17	0.89
EL5	0.81
EN5	0.72
Service	0.86
Rating	0.58

The scree plot of the extracted factors is presented in Figure 12, 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 SEL “Service” and the EL items within SEL Enterprise. This supports including the EN items in the calculation of SEL Enterprise scores.

Figure 12: Scree Plot of the Extracted Factors



Relationship of STAR Early Literacy Enterprise Scores to Common Core State Standards Skills Ratings

As was done in the original STAR Early Literacy Validation Study, in order to have an independent common measure of literacy skills, Renaissance Learning constructed a ten item checklist for teachers to use during the Equivalence Study to rate their students on a wide range of competencies related to developing literacy skills. In keeping with current developments in assessment in the U.S., the competencies to be rated represented key skills in the CCSS developed by the National Governors Association and the Council of Chief State School Officers. As before, 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 Equivalence Study, and the relationship between student skills ratings on the instrument and their scores on STAR Early Literacy Enterprise.

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. This rating instrument was intended to specify a sequence of skills that the teacher could quickly assess for each student, chosen such that a student who can correctly perform the nth skill in the list can almost certainly perform all of the preceding ones correctly as well. Such a list, even though quite short, would enable us reliably to sort students from pre-kindergarten through third grade into an ordered set of skill categories.

A list of ten skill-related items was assembled. Each participating teacher was asked to rate his or her STAR Early Literacy Enterprise Equivalence Study 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 Enterprise Equivalence Research Study 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)

    Sample paragraph:

    Richard likes two things about picking apples. He gets to climb a ladder. He can eat the apples he picks.

    Why does Richard like picking apples?

    A. He likes to be outside.

    B. He likes eating apples.

    C. He likes helping his dad.

Psychometric Properties of the CCSS Skills Ratings

The rating worksheet was scored for each student by assigning one point for each performance task marked by the teacher. The range of possible scores was 0 to 10. Teachers completed skills ratings for 6,708 of the students in the Enterprise Equivalence Study. Table 44 lists data about the psychometric properties of the ten item rating scale, overall and by grade, including the correlations between skills ratings and Scaled Scores. The internal consistency reliability of the skills ratings was 0.84, as estimated by coefficient alpha.

Table 44: Teachers' Ratings of Their Students' Common Core State Standards-Aligned Skills, by Grade

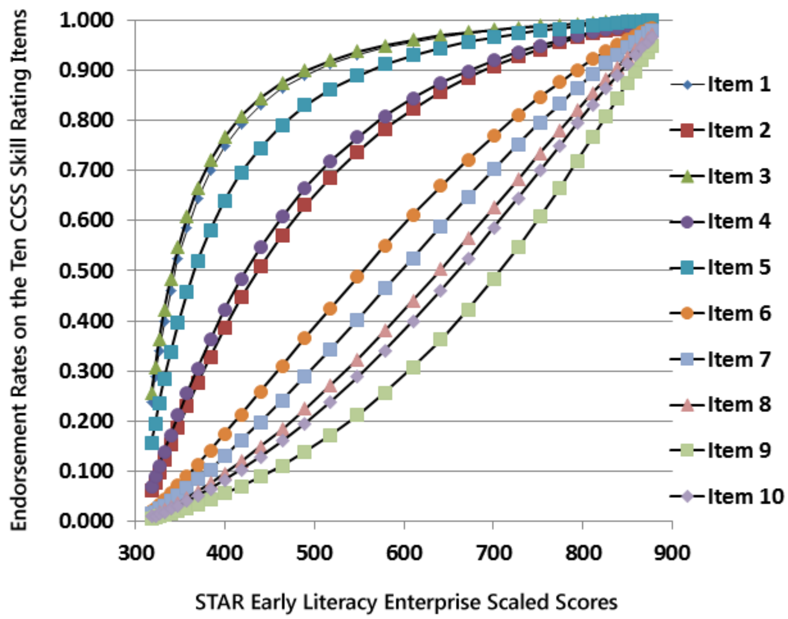
Grade	N	Ratings		Scaled Scores		Correlation of Skills Ratings with Scaled Scores
		Mean	Standard Deviation	Mean	Standard Deviation	
K	2,039	7.5	2.4	687	99	0.47
1	2,552	8.7	2.0	749	98	0.58
2	1,243	9.0	1.9	818	69	0.46
3	874	9.5	1.3	839	59	0.60

Relationship of Scaled Scores to Skills Ratings

As the data in Table 44 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 STAR Early Literacy Enterprise 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 Enterprise test and teachers' ratings of their students' CCSS literacy skills. Within-grade correlations ranged from 0.46 to 0.60.

Figure 13 displays the relationships of each of the ten rating scale items to STAR Early Enterprise Literacy 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 13: Relationship of the Endorsement Rates on the Ten CCSS Skill Rating Items to STAR Early Literacy Enterprise Scaled Scores



As was done in the original STAR Early Literacy Validation Study, in order to have an independent common measure of literacy skills, Renaissance Learning constructed a ten-item checklist for teachers to use during the Equivalence Study to rate their students on a wide range of competencies related to developing literacy skills. In keeping with current developments in assessment in the U.S., the competencies to be rated represented key skills in the Common Core State Standards (CCSS) developed by the National Governors Association and the Council of Chief State School Officers. As before, 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 data on the relationship between student's skills ratings on the instrument and their scores on STAR Early Literacy Enterprise.

To gather the ratings of literacy skills from teachers, a short list of dichotomous items that represent a hierarchy of skills aligned to the CCSS was constructed. This rating instrument was intended to specify a sequence of skills that the teacher could quickly assess for each student, chosen such that a student who can correctly perform the nth skill in the list can almost certainly perform all of the preceding ones correctly as well. Such a list, even though quite short, would enable us reliably to sort students from pre-kindergarten through third grade into an ordered set of skill categories.

A list of ten skill-related items was assembled. Each participating teacher was asked to rate his or her STAR Early Literacy Equivalence Study students as to whether or not they had mastered 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 on page 96.

Table 45 displays summary statistics on the ratings, by grade, along with the mean STAR Early Literacy Enterprise scale scores. The rightmost column contains correlations of the STAR Early Literacy Enterprise scores with the ratings, by grade and for all grades combined.

Table 45: 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

As the data in Table 45 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 STAR Early Literacy Enterprise 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 Enterprise test and teachers' ratings of their students' CCSS literacy skills. Within-grade correlations ranged from 0.46 to 0.60.

Norming

STAR Early Literacy Enterprise 2014 Norming

Nationally representative test score norms were computed for the first time for STAR Early Literacy Enterprise assessments, for introduction at the beginning of the 2014–15 school year. This chapter describes the 2014 norming of STAR Early Literacy Enterprise.

In addition to Scaled Score norms, which are distributions of the scores themselves, Renaissance Learning has developed growth norms for STAR Early Literacy Enterprise. This chapter includes two sections. The first one deals with the 2014 development of the STAR Early Literacy Enterprise test score norms. A second section describes the development and use of the growth norms. Growth norms are very different from test score norms, having different meaning and different uses. Users interested in growth norms should familiarize themselves with the differences, which are made clear in the growth norms section.

Development of Norms for STAR Early Literacy Test Scores

Sample Characteristics

Students' STAR Early Literacy Enterprise data that was available in the Renaissance Place hosted learning environment from fall 2012 to spring 2013 were used for the 2014 STAR Early Literacy Enterprise norming study. The norming sample included students from all US states and the District of Columbia. Information about school and district demographic data were obtained from Market Data Retrieval (MDR), National Center for Education Statistics (NCES), and the US Bureau of Census. Students' demographic data, when recorded by the schools, also included gender, race/ethnicity, Students with Disabilities (SWD), and English Language Learners (ELL).

To obtain a representative sample of the US school population for appropriate fall and spring norming, the first step identified a matched sample of 332,392 students with both fall and spring STAR Early Literacy Enterprise assessment scores. The matched sample of students had completed a STAR Early Literacy assessment in the first three months of the 2012–2013 school year (fall) and also a STAR Early Literacy assessment in the last three months of the 2012–2013 school year (spring). This step insured that the norming process would apply to the same group of students for the full school year irrespective of the time of administration of the STAR Early Literacy Enterprise assessment.

The second step in the norming process was to randomly select at each grade an equal-sized sample of students from each of ten deciles of student achievement performance; this sample of 134,830 students, each with a fall and a spring test record, constituted the final norming sample. To avoid any potential bias in the sample selection, the spring decile of student performance for the matched sample was used in creating the ten decile student achievement performance groups. This step insured that the norming process was appropriate for students within each of ten decile achievement performance groups and to reduce the effects of sample selection bias.

The third step in the norming process was a post-stratification weighting procedure to ensure that the randomized matched student's grade and decile norming sample was adjusted to match a US nationally representative sample. There were three key sample stratification variables used in the norming process: geographic region (Northeast, Southeast, Midwest, and West), socioeconomic status (low SES, below-median SES, above-median SES, and high SES), and school size (< 200 students, 200–499 students, and 500+ students).

The post-stratification process allowed sample adjustments with all possible combinations of three key stratification variables for geographic regions (4 groups) socioeconomic status (4 groups) and school size (3 groups) for a total of 48 combinations of the stratification groups. Specific post-stratification weights were specified for each of the 48 combinations of the stratification groupings to allow the stratification grouping to match the US national population percentages.

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.

Grades. The STAR Early Literacy Enterprise 2014 norming sample included students from grades K–3.

Deciles. Students' STAR Early Literacy Enterprise Scale Scores within each grade were grouped into 10 deciles using the spring 2013 assessment Scale Scores, and then students were randomly sampled from each of the ten achievement decile classifications within each grade level.

Tables 46–48 summarize the key norming variables for the STAR Early Literacy Enterprise norming sample of 134,830 students selected from the norming population of 332,392 students.

Table 46: Sample Characteristics, STAR Early Literacy Enterprise Norming Study 2014 (N= 134,830 Students)

		Students	
		National %	**Sample %**
Geographic Region	Northeast	16.2%	12.8%
	Midwest	21.4%	21.3%
	Southeast	26.8%	27.2%
	West	35.6%	38.7%

Table 46: Sample Characteristics, STAR Early Literacy Enterprise Norming Study 2014 (N= 134,830 Students) (Continued)

		Students	
		National %	**Sample %**
District Socioeconomic Status	Low SES	20.1%	35.1%
	Below-Median SES	26.6%	22.2%
	Above-Median SES	28.9%	29.4%
	High SES	24.0%	13.3%
School Size	< 200 Students	13.0%	3.4%
	200–499 Students	45.0%	41.7%
	500+ Students	41.8%	55.0%
School Metro Code	Rural	25.4%	28.5%
	Suburban	34.0%	26.6%
	Town	11.6%	16.0%
	Urban	28.9%	30.0%

Table 47: Non-Public Schools, STAR Early Literacy Enterprise Norming Study 2014 (N = 134,830 Students)

		Students	
		National %	**Sample %**
School Type	Public	89.3%	98.5%
	Non-Public	10.7%	1.5%
	Catholic	4.2%	1.0%[a]
	Private	6.4%	0.5%[a]

a. One percent of the norming sample represented Catholic schools; 0.5 percent of the norming sample represented private schools. The addition of the Catholic and private schools equals the 1.5% of the sample for non-public schools.

Table 48: Gender and Ethnic Group Participation, STAR Early Literacy Enterprise Norming Study 2014 (N = 134,830 Students)

		Students	
		National %	**Sample %**
Ethnic Group	Asian	4.7%	2.2%
	Black	15.8%	24.0%
	Hispanic	23.8%	27.6%
	Native American	1.5%	1.8%
	White	51.7%	44.6%
	Multiple Ethnicity	2.6%	Not Coded for Analysis
	Not Recorded		55.7%[a]
Gender	Female	50.8%	46.6%
	Male	49.2%	53.4%
	Not Recorded		19.5%[a]

a. Ethnic Group was not recorded in 55.7% of the data records; 19.5% did not have student Gender recorded. Percentages in the table above are based only on those cases with non-missing data for those variables.

The STAR Early Literacy Enterprise 2014 norming sample included 2,819 (2.1%) cases identified as students with disabilities (SWD), and 3,639 (2.7%) identified as English Language Learners (ELL). An estimated 5.8% of the students in the norming sample were gifted and talented (G&T) based on the 2011–2012 school data collected by the Office of Civil Rights (OCR). OCR is a subsidiary of the US Department of Education.

Data Analysis

As noted above, the first step in the norming process was to select a matched sample of students within grades K–3 who had taken a STAR Early Literacy Enterprise assessment in the first three months of 2012–2013 school year and a STAR Early Literacy Enterprise assessment in the last three months of the 2012–2013 school year.

From the matched student sample, the second step in the norming process was to select for each grade a random equal-sized sample from each of the ten deciles of achievement performance. The decile for the spring 2013 STAR Early Literacy Enterprise Scaled Score was used for the decile classification to avoid any sample selection bias. The sample size for each decile within each grade was determined by the sample size available for the smallest decile sample size available per grade.

The third step in the norming process was to employ post-stratification sample weights to make the weighted data closely approximate the distributions of test scores within the national student population. The sample weights were applied to all 48 combinations of stratification variables. Weighted scores were used to compute the norms for both fall and spring at each grade level.

Table 49 shows the fall and spring Scaled Score summary statistics by grade.

Table 49: Summary Statistics on Fall and Spring Weighted Scaled Scores, STAR Early Literacy Enterprise Norming Study—2014 (N = 134,830 Students)

Grade	Sample Size	Fall Scores			Spring Scores		
		Scaled Score Means	Scaled Score Standard Deviations	Scaled Score Medians	Scaled Score Means	Scaled Score Standard Deviations	Scaled Score Medians
K	39,000	523	95	522	610	101	611
1	66,840	635	106	631	738	90	754
2	25,730	720	104	743	788	81	810
3	3,260	766	101	803	818	70	838

The sample sizes are identical for the fall and spring scores within each grade in Table 49 since students were selected for the norming sample if there were matched fall and spring scores from the same group of students.

The norm-referenced scores are determined from both the fall and spring testing periods used for the norming. Table 65 (see page 157) presents an abridged version of the Scaled Score to Percentile Rank conversion tables for STAR Early Literacy Enterprise. The actual table in the software includes data for each of the monthly grade placement values 0.0–2.9. Since the norming of STAR Early Literacy Enterprise occurred in the months of August, September, and October of the school year for the fall testing period and April, May, and June of the school year for the spring testing period, empirical values were thus established for the fall and spring norming periods. The data for the remaining monthly values were established by interpolation between the two empirical testing periods. Table 65 (see page 157) presents Scaled Score to Percentile Rank conversion by grade (at month 7 of the school year) as an abridgment of the larger table included in the software. The expanded table provided in the software provides normative information that is most relevant for the specific time period in which each student takes the STAR Early Literacy Enterprise assessment.

Grade Equivalent (GE) scores for each grade and each month of the school year were computed by interpolation between the median fall and spring Scaled Scores for each grade. The interpolated values for the Grade Equivalent values were smoothed using accepted statistical procedures. The Scaled Score to Grade Equivalent conversion table is presented in Table 50.

Table 50: STAR Early Literacy Enterprise Scaled Score to Grade Equivalent Conversions

Scaled Score	Grade Equivalent	Scaled Score	Grade Equivalent
300–520	0	747–754	2
521–536	0.1	755–761	2.1
537–545	0.2	762–767	2.2
546–553	0.3	768–773	2.3
554–560	0.4	774–779	2.4
561–569	0.5	780–785	2.5
570–579	0.6	786–790	2.6
580–590	0.7	791–795	2.7
591–603	0.8	796–799	2.8
604–618	0.9	800–804	2.9
619–633	1	805–808	3
634–648	1.1	809–812	3.1
649–664	1.2	813–815	3.2
665–678	1.3	816–819	3.3
679–692	1.4	820–822	3.4
693–705	1.5	823–825	3.5
706–717	1.6	826–828	3.6
718–728	1.7	829–831	3.7
729–737	1.8	832–835	3.8
738–746	1.9	836–900	3.9

Growth Norms

To enhance the utility of STAR assessments for indexing growth, two types of growth metrics are calculated annually: Student Growth Percentile (SGP) and growth norms. Both are norm-referenced estimates that compare a student's growth to that of his or her academic peers nationwide. SGP 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 starting scores. SGPs range from 1–99 and are interpreted similar to Percentile Ranks. Growth norms are the median scaled score change observed for students within a given grade and pre-test decile, and thus facilitate norm-referenced comparisons of student absolute growth. Both SGPs and growth norms can be useful for setting realistic goals and gauging whether a student's growth is typical.

At present, the growth norms in STAR Early Literacy are based on student assessments (N = 1,012,475). Growth norms provide a reference to distributions of student growth over time and across the academic year. Growth norms were developed to index growth of student groups from different grades and with different levels of initial performance on STAR Early Literacy. This provides a method of comparing a student's observed growth over a period of time to growth made by students of a similar grade and achievement level.

Students develop at different rates within each grade and depending on where they score in the overall distribution of performance, students who score in the top decile for a grade do not, and should not be expected to, grow at the same rate across the academic year as students in the middle or lower deciles, and vice versa. Growth rates of students should be compared to students of similar academic achievement levels; otherwise, there is the potential for inappropriately expecting too much or too little growth from certain students.

Growth norms were developed by following students across the entire academic year. Students were tested both at the beginning and end of the school year. To normalize differences in time between the initial and final test, change in score from fall to spring testing was divided by the number of weeks between the assessments to obtain the rate of growth per week.

Within each grade, students were divided into decile groups based on their percentile ranks on the initial STAR Early Literacy test of the school year, resulting in 10 decile groups for each grade. For each decile within each grade, the median weekly scaled score change was computed.

Using data retrieved from the hosted Renaissance Place customer database, growth norms are updated annually to reflect changes in educational practices, and ensure students' observed growth is being referenced against an up-to-date student group.

Score Distributions

Scaled Scores: Score Distributions

Non-adaptive Calibration Study Data. At the completion of the item calibration process, the resulting item parameters were applied to the item response data of the 246 calibration forms to calculate Rasch ability estimates for all students, as well as Scaled Scores, Domain Scores, and Skill Scores. Table 51 contains Scaled Score summary statistics by grade, including means, standard deviations, numbers of observations, and 10th, 50th and 90th percentile values.

Table 51: Distributional Statistics of Scaled Scores in the STAR Early Literacy Calibration Study Sample, by Grade

				Percentile		
Grade	**N**	**Mean**	**SD**	**10**	**50**	**90**
Pre-Kindergarten	3,335	518	87	412	511	625
Kindergarten	8,453	585	85	482	580	702
Grade 1	15,759	701	83	591	703	810
Grade 2	9,959	763	82	647	779	856
Grade 3	8,920	812	63	734	826	873

Adaptive Validation Study Data. Table 52 contains Scaled Score summary statistics from the adaptive Validation Study by grade, separately for the US and Canadian samples. Like Table 51, Table 52 contains means and standard deviations. It also provides more detailed percentile data, ranging from the 1st to the 99th percentiles.

Table 52: Scaled Score Distributions of US and Canadian Students in the STAR Early Literacy Validation Study Sample, by Grade

				Percentile										
Grade	**N**	**Mean**	**SD**	**1**	**5**	**10**	**15**	**25**	**50**	**75**	**85**	**90**	**95**	**99**
US														
Pre-K	449	472	112	345	359	368	377	393	439	511	560	621	711	850
K	1,982	588	103	370	420	450	477	515	585	659	701	729	766	815
1	2,423	742	94	450	552	608	643	693	763	812	833	843	857	879
2	1,769	796	79	486	640	697	727	766	816	848	863	870	882	890
3	1,905	823	63	592	702	748	772	801	841	866	875	881	885	891

Table 52: Scaled Score Distributions of US and Canadian Students in the STAR Early Literacy Validation Study Sample, by Grade (Continued)

				Percentile										
Grade	**N**	**Mean**	**SD**	**1**	**5**	**10**	**15**	**25**	**50**	**75**	**85**	**90**	**95**	**99**
Canada														
Pre-K	78	469	75	350	369	396	400	418	467	508	538	567	610	780
K	325	570	107	371	410	427	450	492	564	640	689	712	758	873
1	400	677	112	406	458	513	556	610	688	765	794	813	836	858
2	371	782	77	541	614	672	707	748	800	839	852	861	871	883
3	359	815	71	587	665	734	760	789	834	860	871	878	884	891

Literacy Classification: Score Distributions

Tables 53 and 54 list the frequency of the three literacy classifications by grade, for the US and Canadian samples, respectively. These tables include row-wise percentages that indicate the relative distributions of the three categories within grade.

Table 53: Distribution of the Literacy Classification by Grade for the US Sample in the Validation Study

	Literacy Classification			
Grade	**Emergent Reader**	**Transitional Reader**	**Probable Reader**	**Total**
Pre-K	414 (92%)	16 (4%)	19 (4%)	449
K	1,571 (79%)	325 (16%)	86 (4%)	1,982
1	497 (21%)	845 (35%)	1,081 (45%)	2,423
2	135 (8%)	355 (20%)	1,279 (72%)	1,769
3	69 (4%)	226 (12%)	1,610 (85%)	1,905

Table 54: Distribution of the Literacy Classification by Grade for the Canadian Sample in the Validation Study

	Literacy Classification			
Grade	**Emergent Reader**	**Transitional Reader**	**Probable Reader**	**Total**
Pre-K	77 (99%)	0 (0%)	1 (1%)	78
K	271 (83%)	42 (13%)	12 (4%)	325
1	182 (46%)	138 (34%)	80 (20%)	400
2	42 (11%)	91 (25%)	238 (64%)	371
3	21 (6%)	48 (13%)	290 (81%)	359

Score Definitions

For its internal computations, STAR Early Literacy Enterprise 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 Enterprise software uses a proprietary Maximum-Likelihood IRT estimation procedure to avoid any potential bias in the Scaled Scores. All STAR Early Literacy Enterprise 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, proficiency scores, and Estimated Oral Reading Fluency scores (Est. ORF). On the basis of Scaled Scores, students taking STAR Early Literacy Enterprise are categorized into one of three literacy classifications (see page 111). In addition, STAR Early Literacy Enterprise uses two types of proficiency scores: Sub-domain Scores and Skill Set Scores.

The four sections that follow present score definitions, followed by score distribution summary data from the STAR Early Literacy non-adaptive Calibration Study and the adaptive Validation Research Study.

Scaled Scores

Scaled Scores are the fundamental scores used to summarize students' performance on STAR Early Literacy Enterprise tests. Upon completion of STAR Early Literacy Enterprise, each student receives a single-valued Scaled Score. The Scaled Score is a non-linear, monotonic transformation of the Rasch ability estimate resulting from the adaptive test. STAR Early Literacy Enterprise Scaled Scores range from 300 to 900.

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 Enterprise 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.

Sub-domain and Skill Set Scores

STAR Early Literacy Enterprise 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 Enterprise item bank. These proficiency scores are referred to in STAR Early Literacy Enterprise 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 Enterprise 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 Enterprise items in a specific skill 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 is used to express each Sub-domain or Skill Set Score.

Literacy Classification

STAR Early Literacy Enterprise 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 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 Enterprise item bank. Similarly, to be classified as a Probable Reader, mastery of higher-level skills must be apparent. A detailed rationale for the choice of 675 and 775 as cut scores for this three-part classification is presented in "STAR Early Literacy Enterprise in the Classroom" on page 114.

Estimated Oral Reading Fluency (Est. ORF)

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

Est. ORF is reported as 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.

The Est. ORF score was computed using the results of a large-scale research study investigating the linkage between estimates of oral reading fluency[16] and both STAR Early Literacy and STAR Reading scores. An equipercentile linking was done between STAR Reading scores and oral reading fluency providing an estimate of the oral reading fluency for each scale score unit on STAR Reading for grades 1–4 independently. A linear equating between the STAR Early Literacy and STAR Reading score scales was also undertaken. STAR Early Literacy's estimates of oral reading fluency are derived by first transforming the STAR Early Literacy scale score to an equivalent STAR Reading score, then looking up the corresponding estimated oral reading fluency score. There are separate tables of corresponding STAR Reading-to-oral reading fluency scores for each grade from 1–4; however, STAR Early Literacy reports estimated oral reading fluency only for grades 1–3.

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. SGPs provide a measure of how a student changed from one STAR testing window[17] to the next relative to other students with similar starting STAR Early Literacy Enterprise 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

16. The research study is described in the Validity section of this technical manual. See "Concurrent Validity of Estimated Oral Reading Score" on page 77. Additional details are presented in the *STAR Reading Technical Manual*.
17. We collect data for our growth norms during three different time periods: fall, winter, and spring. More information about these time periods is provided on page 143.

70% of students nationwide in the same grade with a similar beginning (pretest) STAR Early Literacy Enterprise score. 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 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 Enterprise 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. SGP 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.

STAR Early Literacy Enterprise in the Classroom

Recommended Uses

Intended Population

STAR Early Literacy Enterprise was designed for regular assessment of literacy skills and concepts. Although intended primarily for use from pre-kindergarten through grade 2, it may be used to assess any student who is not yet an independent reader. Because students vary widely in age at the pre-kindergarten level, teachers should exercise discretion when using STAR Early Literacy Enterprise with this population. In the research and development of STAR Early Literacy Enterprise, children three and four years of age took the test and most attained Scaled Scores well above the minimum level. Because successful test administration requires the ability to use the mouse or keyboard to select answers to test questions, children who are not successful in the hands-on exercise section of STAR Early Literacy Enterprise should not be encouraged to take the assessment section.

STAR Early Literacy Enterprise may also be used for assessment of grade 3 students, and of remedial students in grades 4 and above. By the end of grade 3, most students will have mastered the literacy skills that STAR Early Literacy Enterprise assesses. Such students will achieve Scaled Scores approaching 900, the top end of the scale.[18] Such high STAR Early Literacy Enterprise scores will be useful for determining that a student is functioning at a mastery level. However, STAR Early Literacy Enterprise may not be a useful measure of growth among students who have already attained mastery of literacy skills. Beyond that point, an assessment of traditional reading growth is more appropriate, and teachers may choose to administer a standardized reading achievement test, such as STAR Reading.

In terms of administering STAR Early Literacy Enterprise to remedial students at grade 3 and beyond, teachers should recall that items for the test were designed specifically for young children. In making the decision whether to use STAR Early Literacy Enterprise to assess older students, teachers should evaluate whether the format and content of the items are appropriate for their students. While some older students may find the STAR Early Literacy Enterprise items engaging, others may not be motivated to perform their best on a test that seems designed for a younger age group.

18. Perfect scores of 900 are rare among third graders; scores of 850 and above indicate a substantial degree of mastery of early literacy skills.

Uses

STAR Early Literacy Enterprise provides teachers with immediate feedback that highlights instructional needs and enables teachers to target literacy instruction in order to improve the overall literacy skills of their students by some measurable means.

STAR Early Literacy Enterprise was developed as a criterion-referenced assessment system. Students are compared to a criterion or a standard and an absolute score is reported. The norming study of summer 2014 enhanced the product to include relative scores to compare students to one another.

Using a criterion-referenced assessment system has many advantages for teachers. First, STAR Early Literacy Enterprise's skills-based scores can be used by teachers to guide planning and instruction. In addition, teachers may administer STAR Early Literacy Enterprise repeatedly, allowing for ongoing monitoring of student progress. As a result, teachers will find STAR Early Literacy Enterprise useful for the following tasks:

- Literacy classification
- Individual readiness screening
- Match books to early readers
- Overall early literacy assessment
- Component literacy skills assessment
- Growth measurement
- Progress monitoring

Approaches and Rationales for Recommended Uses

Literacy Classification

STAR Early Literacy Scaled Scores are used to classify every student into one of three broad stages of reading development:

- Emergent Reader: Scaled Scores ranging from 300 to 674.
  - Early Emergent Reader: Scaled Scores ranging from 300 to 487.
  - Late Emergent Reader: Scaled Scores ranging from 488 to 674.
- Transitional Reader: Scaled Scores from 675 to 774.
- Probable Reader: Scaled Scores from 775 to 900.

The rationale for the choice of 675 and 775 as cutoff scores was based on the relationships between Scaled Scores and proficiency in the sub-domains and discrete skills identified in “Content and Item Development” on page 15. Specifically, the Calibration Study data showed that students with Scaled Scores of 675 and above have Skill Set Scores above the 80 percent mastery level in 5 sets of skills that are critical to beginning reading, particularly alphabet skills. Table 55 on the next page lists the five relevant skill sets, which include one set of skills in the Visual Discrimination sub-domain, two sets of skills in the of Early Numeracy sub-domain, and two sets of skills in the sets in the Alphabetic Principle sub-domain.

Table 55: Critical Skills for Differentiating Emergent Readers from Transitional Readers

Sub-Domain	Skill Set	Skill
Visual Discrimination	Letters	Differentiate lowercase letters
		Differentiate uppercase letters
		Differentiate lowercase letters in mixed set
		Differentiate uppercase letters in mixed set
Early Numeracy	Number Naming and Number Identification	Recognize numbers 0–20
	Number Object Correspondence	Count 1–20
		Recognize ordinal numbers 1st–10th
Alphabetic Principle	Alphabetic Knowledge	Recognize lowercase letters
		Recognize uppercase letters
		Match lowercase with uppercase letters
		Match uppercase with lowercase letters
		Distinguish numbers from letters
	Letter Sounds	Recognize sounds of lowercase letters
		Recognize sounds of uppercase letters

Further along the developmental scale, students with scores above 775 are estimated to be able to answer at least 80 percent of all items in the STAR Early Literacy Enterprise item bank correctly, with a mastery of 70 percent or better in all ten literacy sub-domains. Students classified as “Probable Readers” are likely to be successful in taking a STAR Reading test.

Because the Emergent Reader includes a very broad range of skill levels, this classification is divided into Early Emergent Reader (Scaled Scores from 300 to 487) and Late Emergent Reader (Scaled Scores from 488 to 674). Students at the early Emergent Reader stage are beginning to understand that printed text has meaning. They are learning that reading involves printed words and sentences, and that print flows from left to right and from the top to the bottom of the page.

They are also beginning to identify colors, shapes, numbers, and letters. Early Emergent readers can relate pictures to words. These students are beginning to recognize and tell the difference between words and letters. They can probably relate some letters with their sounds and are likely beginning to separate spoken words into individual units of sound.

At the Late Emergent Reader stage, students can identify most of the letters of the alphabet and can match most of the letters to their sounds. They are beginning to "read" picture books and familiar words around their home. Through repeated reading of favorite books with an adult, children at this stage are building their vocabularies, listening skills, and understanding of print. Late Emergent Readers can recognize some printed words in their surroundings, including signs and their names. They are also learning to separate spoken words into smaller parts, such as m- and -at for "mat." Late Emergent Readers are probably beginning to "sound out" simple printed words and are starting to get meaning from text with their growing knowledge of letter sounds and word structure.

Students at the Transitional Reader stage have mastered their alphabet skills and letter-sound relationships. They can identify many beginning and ending consonant sounds and long and short vowel sounds. Students at this stage are most likely able to blend sounds and word parts to read simple words. They are likely using a variety of strategies to figure out words, such as pictures, story patterns, and phonics. Transitional Readers are generally starting to apply basic concepts about print and books to unfamiliar text. Students at this stage are beginning to read unfamiliar words and easy-reader material, but are not yet fluent, independent readers.

Students at the Probable Reader stage are becoming proficient at recognizing many words, both in and out of context. They spend less time identifying and sounding out words and more time understanding what they have read. They can blend sounds and word parts to read words and sentences more quickly, smoothly, and independently. Probable Readers are starting to challenge themselves with longer picture and chapter books. They are increasingly able to select books that interest them, to monitor their own reading, and to self-correct as needed. Probable Readers are generally able to read silently and to read aloud some easy texts with accuracy, fluency, and expression.

Screening Assessment

STAR Early Literacy Enterprise can be used in several ways as a screening instrument as recommended by No Child Left Behind. The purpose of screening is to identify children who are at risk for delayed development or academic failure and who require additional reading instruction or further diagnosis. STAR Early Literacy Enterprise can also be used for this further diagnosis.

Benchmarks and Cut Scores

Screening applications of STAR Early Literacy Enterprise can be done within the context of benchmarks and cut scores. These scores help educators identify which students require some form of intervention to accelerate growth and move toward proficiency.

Benchmarks are the minimum performance levels students are expected to reach by certain points of the year in order to meet end-of-year performance goals. The end-of-year benchmark typically represents the minimum level of performance required by state or local standards. Benchmarks are always grade specific, e.g., the 2nd grade benchmark. In Table 56, the 40th and 50th percentile represent two benchmark options. Schools should select one based on their state recommendations or local guidelines. Experts often recommend the grade-level benchmark be set at the 40th percentile. Default benchmarks in STAR Early Literacy are set at the 40th percentile. Transition benchmarks, for educators who choose to use them, are higher in grades after kindergarten. When students become probable readers they should begin taking STAR Reading. When administering STAR Early Literacy to students after kindergarten, benchmark percentiles should be at the 55th (1st grade), 70th (2nd grade), and 80th (3rd grade). To understand this increased percentile, refer to the technical note on page 120.

A cut score is used to determine which students may need additional assistance to move toward the end of year benchmark. In Table 56, the 10th, 20th, and 25th percentile represent three cut scores options. Schools should select one based on their state recommendations or local guidelines.

Benchmarks and cut scores do not replace educator judgment; they inform it. Proper determination of cut scores is key to successful implementation of intervention and other data-based decision making processes.

Table 56 (on the next page) offers benchmarks and cut scores for the three typical screening periods of the school year—fall (September), winter (January), and spring (May). For example, 1st grade students tested during the winter administration who scored below 645 could be considered to be below benchmark, whereas students scoring at or above 672 could be considered at or above benchmark.

A second method of screening students is the fixed standard approach. In this method, early childhood educators in a school or district choose a STAR Early Literacy Enterprise score that reflects their best judgment about which students are at risk. At the discretion of teachers or local school administrators, standards could be based on attainment of minimum Scaled Scores or specific Sub-domain or Skill Set Scores.

Table 56: Default Benchmarks[a]

		Fall September		Winter January		Spring May		Moderate Growth Rate
Grade	Percentile	Scaled Score	Est. ORF[b]	Scaled Score	Est. ORF[b]	Scaled Score	Est. ORF[b]	Scaled Score/Week
K	10	399		430		469		5.7
	20	437		472		512		5.7
	25	452		489		529		5.6
	40	496		534		573		5.4
	50	522		561		601		5.3
	75	582		626		669		4.1
	90	647		691		732		3.5
1	10	499	0	549	6	603	14	6.7
	20	545	6	601	13	657	20	6.2
	25	561	9	619	15	675	23	5.8
	40	603	14	663	22	718	29	5.4
	50	631	17	690	25	742	35	5.1
	75	713	28	759	41	797	60	3.5
	90	778	50	809	67	833	84	2.9
2	10	566	10	617	17	668	23	5.1
	20	630	19	679	25	724	31	4.1
	25	657	22	702	27	742	35	3.4
	40	707	28	748	36	782	50	3
	50	743	35	776	47	804	60	2.5
	75	805	61	826	75	844	92	1.4
	90	841	89	857	108	869	128	1.2
3	10	608	14	672	24	730	34	3.8
	20	690	27	735	35	774	48	2.5
	25	717	32	756	43	789	52	2
	40	783	51	804	57	821	66	1.7
	50	803	57	821	66	835	75	1.4
	75	841	79	853	93	862	106	0.7
	90	867	113	874	129	879	146	0.5

a. The default STAR Early Literacy benchmarks (in the software) are based on the updated 2014–2015 norms.

b. Est. ORF: Estimated Oral Reading Fluency is only reported for grades 1–3.

For example, a school might use sub-domain score distribution statistics to establish a minimum score of 30 on Phonemic Awareness. They would base this decision on the strong relationship between phonemic awareness and later reading success. Students who do not exceed this minimum score would be provided additional reading instruction.

The third method is to establish local norms using data from prior years. Educators would set standards based on the historic distribution of STAR Early Literacy Enterprise scores within a particular school, district, or other local unit. Using the same process described above, the choice of minimum scores would reflect STAR Early Literacy Enterprise Scaled Scores or specific Sub-domain or Skill Set Scores.

Table 57: Transition Benchmarks

		Fall September		Winter January		Spring May		Moderate Growth Rate
Grade	Percentile[a]	Scaled Score	Est. ORF	Scaled Score	Est. ORF	Scaled Score	Est. ORF	Scaled Score/Week
K	10	399		430		469		5.6
	25	452		489		529		5.4
	40	496		534		573		5.2
1	20	545	6	601	13	657	20	6.0
	40	603	14	663	22	718	29	5.3
	55	647	19	704	26	755	40	4.7
2	40	707	28	748	36	782	50	3.0
	60	771	45	798	57	821	71	2.1
	70	793	55	817	68	836	84	1.8
3	45	794	54	812	61	828	70	1.3
	65	827	69	842	80	853	93	0.8
	80	849	87	860	103	869	117	0.6

a. Urgent Intervention Intervention Benchmark.

Technical Note: Rationale for Changes in the STAR Early Literacy Cut Scores

In previous years, based in part on the advice of national experts in RTI, Renaissance suggested cut scores predicated on students' percentile ranks, for classifying students into four intervention categories: Urgent, Intervention, On-watch, and at or above Benchmark. The three suggested cut scores for this four-fold classification were the 10th, 25th, and 40th percentiles. The same three percentile cut scores were used for all three STAR assessments: Early Literacy, Reading, and Math.

In 2014, two new developments led to reconsideration of the use of the same PRs (10, 25, and 40) for both STAR Early Literacy and for STAR Reading. The first of these was the development of an experimental single score scale which could be used to summarize performance on either STAR Early Literacy or STAR Reading; we call this the STAR Early Reading unified score scale. While STAR scores are not yet being reported on the unified scale, that scale can be used to determine what scale scores on Early Literacy are equivalent to STAR Reading scale scores, and vice versa.

The second development of note was the development of nationally representative norms for STAR Early Literacy. In previous years, STAR Early Literacy did not report percentile ranks because the only norms available for it were based on a self-selected 2001 research study sample. Those norms were the basis for the default percentile rank cut scores (10, 25, and 40) used in the STAR Early Literacy Screening Report, but they were not regarded as representative of the national population.

Those research norms for SEL were useful in RTI applications when no other information about student performance relative to a national sample was available. However, in some instances, students took both STAR Early Literacy and STAR Reading. In such cases, the same students might be placed in notably different intervention categories by the two tests. Often students with relatively high Early Literacy percentile ranks attained much lower percentile ranks on their STAR Reading tests; as a consequence, a number of students might be identified as at or above a benchmark based on their Early Literacy percentiles, but in need of intervention based on STAR Reading percentiles.

Data

The availability in 2014 of the new, nationally representative STAR Early Literacy norms mitigated that somewhat, but not completely. For example, Table 58 displays the number of students classified as above or below the 25th percentile based on the 2014 STAR Early Literacy norms and the 2014 STAR Reading norms.

Table 58: Grade 3 Students Identified as "At-Risk" by STAR Reading and STAR Early Literacy Using Each Test's 25th Percentile as Cut Score

		STAR Early Literacy	
		Below 25th Percentile	**At or Above 25th Percentile**
STAR Reading	**Below 25th Percentile**	4,553	3,304
	At or Above 25th Percentile	367	6,730

In this example of 3rd grade spring test scores, STAR Reading found 7,857 students (4,553 + 3,304) below the 25th percentile; STAR Early Literacy found just 4,920 (4,553 + 367). The overall agreement rate between the two tests was 75%. But clearly, if the 25th percentile is used as the cut point, STAR Early Literacy identified a much smaller percentage of students as at-risk than did STAR Reading—about 37% fewer. Some teachers interpreted this to mean that “STAR Early Literacy is too easy.” But that’s not the case; what’s to blame here is that the norms groups are not comparable, and therefore the 25th percentile of the STAR Early Literacy score distribution is not equivalent to the 25th percentile of the STAR Reading distribution.

This may seem counter-intuitive, since with the 2014 norms both the STAR Early Literacy and STAR Reading percentile ranks are based on nationally representative samples. Therefore, shouldn’t the 25th percentiles of both tests cut off the same number of students? The answer is “yes,” but only if the 2014 norming samples for the two tests included the same or highly similar groups of students.

Should we expect the samples to be comparable? Not if different selection factors worked to create samples that were not comparable. Consider: the 3rd-grade norms for each of the two tests were based on samples of all of the students who took that test both in the fall and the spring of 3rd grade. Yet the 3rd-grade sample size for STAR Reading was over 270,000, while the sample size for STAR Early Literacy was just over 3,000.

Why such a large disparity? Because by 3rd grade, only the least proficient readers tend to take STAR Early Literacy. The same thing tends to happen in every grade where both tests are used, but the disparity between the two groups—the students who take STAR Early Literacy rather than STAR Reading—gets larger with every grade. As the grade increases, the percentage of students taking STAR Early Literacy decreases, and the reading levels of those students fall farther and farther behind those of the students who take STAR Reading. Table 59 illustrates the differences in reading ability between STAR Early Literacy and STAR Reading test-takers, by grade, in the fall and spring of the 2012–2013 school year. All scores are averages of thousands of students. The STAR Early Literacy scale scores have been transformed to equivalent STAR Reading scale scores to make scores on the two tests comparable.

Table 59: Differences in Reading Ability between STAR Early Literacy and STAR Reading Test-Takers (by Grade, in Fall and Spring of 2012–2013 School Year)

Grade	Fall		Spring	
	STAR Early Literacy	STAR Reading	STAR Early Literacy	STAR Reading
K	57	81	79	83
1	74	80	131	159
2	102	167	206	281
3	151	312	213	399

As Table 59 shows, at every grade, both in fall and spring the average scale scores of students who took STAR Reading only are higher than the scores of those who took STAR Early Literacy only. The differences are small in Grade K, but increase with each grade; the differences in grades 1–3 also tend to be larger in spring than in the fall.

Table 60 shows the same thing for grades 1–3, but expressed in terms of the percentile rank equivalent of the average scale scores in Table 59. (There are no percentiles for grade K because there are no STAR Reading norms for Kindergarten.)

Table 60: Differences in Reading Ability between STAR Early Literacy and STAR Reading Test-Takers (Grades 1–3, in Fall and Spring of 2012–2013 School Year), Expressed as Percentile Rank Equivalents

Grade	Fall		Spring	
	STAR Early Literacy	STAR Reading	STAR Early Literacy	STAR Reading
1	38	56	39	48
2	18	34	17	37
3	7	38	5	41

The pattern is the same as for the scale scores: at all three grades, fall and spring alike, the percentile rank equivalents of students who took only STAR Reading are materially higher than for the students who took STAR Early Literacy only. Both the scale score data in Table 59 and the percentile equivalents in Table 60, illustrate that the score distributions of students taking STAR Early Literacy at any of the listed grades were lower than the score distributions of the STAR Reading test-takers. In short, lower-performing students tended to take STAR Early Literacy, and higher-performing students tended to take STAR Reading. Because the score distributions are appreciably different, norms based on STAR Early Literacy test-takers are not comparable to norms for STAR Reading test-takers, even those for students at the same grade.

To achieve better agreement between STAR Early Literacy and STAR Reading in identification of students who are at risk, in need of intervention, etc., it is necessary to use different percentile rank cut scores for the two tests. Analysis suggests that, for 3rd grade students, the 65th STAR Early Literacy percentile is roughly equivalent to STAR Reading's 25th percentile. Table 61 displays the number of students classified at-risk by the two tests, using the 25th and 65th percentiles as cut scores for STAR Reading and STAR Early Literacy, respectively.

Table 61: Grade 3 Students Identified as "At-Risk" by Using STAR Reading's 25th Percentile and STAR Early Literacy's 65th Percentile as Cut Score

		STAR Early Literacy	
		Below 65th Percentile	At or Above 65th Percentile
STAR Reading	Below 25th Percentile	7,290	567
	At or Above 25th Percentile	2,682	4,415

In this example of 3rd grade spring test scores, STAR Reading found 7,857 students (7,290+ 567) below the 25th percentile (just as it had in Table 58). Using the 65th percentile as the cut score, STAR Early Literacy found 9,972 (7,290 + 2,682); the overall agreement rate between the two tests was 78%, compared to 75% in Table 58. In contrast to the results shown in Table 58, STAR Early Literacy identified about 27% more students as at-risk than did STAR Reading.

Conclusion

The data presented above provides evidence of two related facts:

1. The distributions of reading achievement were different for students who took STAR Early Literacy than for those who took STAR Reading Enterprise, for each grade (K–3) and at each season (fall and spring). STAR Reading test-takers scored higher than students who took only STAR Early Literacy, and the differences increased in magnitude with each increasing grade level, and from fall to spring within each grade.

2. The norm-referenced scores—specifically percentile ranks—for the two tests are not comparable. For example, Table 58 showed that many fewer students scored below the 25th percentile on STAR Early Literacy than on STAR Reading, holding grade and season constant. Table 61 demonstrated that if

STAR Early Literacy and STAR Reading are expected to identify similar numbers of students at risk, the norm-referenced cut score (percentile rank) must be much higher for STAR Early Literacy than for STAR Reading. How much higher varies from one grade to the next, and from one season to another.

Placement Screening

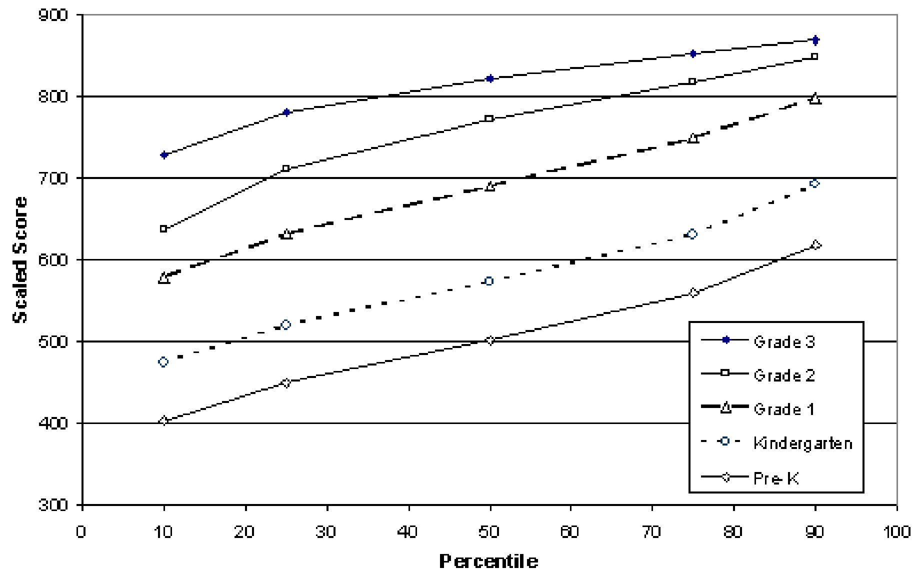
Under No Child Left Behind, the term “screening assessment” is used to describe the process of identifying students who are at risk of reading failure. There is also another approach to screening, using an assessment to determine the education level in which a student should be placed.

Placement screening is typically used to determine if children who are at the younger end of their age cohort should be placed in pre-kindergarten or kindergarten. The typical age at which children begin kindergarten is five. Children who are born late in the year, from October through December, may be 10 months younger than some of their peers. Although this sounds like a short period of time, it is more than fifteen percent of the child’s lifetime. Moreover, it is during these months that critical early literacy development takes place.

STAR Early Literacy Enterprise can provide information that will help parents and educators make the correct placement decision. Using best judgment or historic local norms, those involved in the child’s education can decide on the Scaled Score or specific Sub-domain or Skill Set Scores that can serve as the indicator for placement into pre-kindergarten or kindergarten. By using threshold scores, parents and educators are more likely to make the decision that increases the likelihood that the child will have a positive educational experience.

Figure 14 illustrates a graphical approach to displaying local norms for grades pre-kindergarten through 3 in a hypothetical school district. To ensure that affected students will be capable of progressing along with older class members, teachers and administrators should establish qualifying scores with care. To this end, setting a minimum score between the 25th and 50th percentiles of the reference group is suggested.

Figure 14: A Graphical Example of Local Norms: Percentiles of STAR Early Literacy Enterprise Scores by Grade, Adjusted to the Beginning of Each School Year



Match Early Readers with Books

To help teachers match books to early readers, Table 62 displays correspondences between STAR Early Literacy Scaled Scores and literacy development classifications on the one hand, and STAR Reading Scaled Scores, GE scores, and ZPD ranges on the other.

This table was developed by linking (equating) STAR Early Literacy and STAR Reading Rasch ability scores, then converting the linkage results to scaled scores of the two assessments.

The linking analysis began by identifying students who took the Enterprise versions of both STAR Early Literacy and STAR Reading concurrently during the 2012–2013 school year. More than 300,000 students who had taken both tests within 15 days of one another formed the concurrent tests data set. The linkage work employed linear equating analyses of a random sample of 25,000 matched pairs of scores from that data set. STAR Early Literacy Enterprise and STAR Reading measure related constructs, and were highly correlated ($r = 0.71$) in the concurrent data set.

STAR Reading GE scores in Table 62 are based on 2014 norming data for the Enterprise version of the test, and therefore the scale score ranges corresponding to the GE scores have changed somewhat compared to earlier published versions of the table. STAR Reading ZPD ranges are determined by the GE scores; the correspondence between GE scores and ZPD ranges is unrelated to the new norms, and has not changed.

The primary purpose of Table 62 is to provide a link between STAR Early Literacy scale scores and STAR Reading GE scores, and thus to ZPD ranges, in order to facilitate matching readers to books in Accelerated Reader. It can also be used to identify STAR Early Literacy and STAR Reading scale scores that are approximately equivalent. However, those tests are not the same test, and do measure somewhat different aspects of reading development. Therefore, Table 62 is meant only as a general guide. It should not be used to make decisions about students' early reading achievement or to calculate actual GE scores.

Table 62: Relating STAR Early Literacy Enterprise Scores to STAR Reading Scores

STAR Early Literacy Enterprise		STAR Reading			
Scale Score Range	**Literacy Classification**	**Scale Score Range**	**GE**	**ZPD Range**	**Recommended Assessment(s)**
300–382	Emergent Reader	NA	NA	NA	STAR Early Literacy Enterprise
383–393		0–6	0.0	0.0–1.0	
394–396		7–8	0.1	0.1–1.1	
397–418		9–15	0.2	0.2–1.2	
419–422		16–21	0.3	0.3–1.3	
423–439		22–28	0.4	0.4–1.4	
440–456		29–35	0.5	0.5–1.5	
457–475		36–42	0.6	0.6–1.6	
476–495		43–49	0.7	0.7–1.7	
496–513		50–55	0.8	0.8–1.8	
514–555		56–62	0.9	0.9–1.9	
556–594		63–68	1.0	1.0–2.0	
595–628		69–73	1.1	1.1–2.1	
629–674		74–81	1.2	1.2–2.2	
675–720	Transitional Reader SEL SS = 675	82–92	1.3	1.3–2.3	
721–743		93–105	1.4	1.4–2.4	
744–756		106–120	1.5	1.5–2.5	STAR Early Literacy Enterprise and STAR Reading
757–766		121–137	1.6	1.6–2.6	
767–776	Probable Reader SEL SS = 775	138–153	1.7	1.7–2.7	
777–787		154–171	1.8	1.8–2.8	
788–797		172–188	1.9	1.9–2.9	STAR Reading
798–806		189–206	2.0	2.0–3.0	
807–815		207–223	2.1	2.1–3.1	
816–823		224–240	2.2	2.1–3.1	
824–830		241–257	2.3	2.2–3.2	
831–836		258–273	2.4	2.2–3.2	

Table 62: Relating STAR Early Literacy Enterprise Scores to STAR Reading Scores (Continued)

STAR Early Literacy Enterprise		STAR Reading			Recommended Assessment(s)
Scale Score Range	Literacy Classification	Scale Score Range	GE	ZPD Range	
837–841	Probable Reader (continued)	274–288	2.5	2.3–3.3	STAR Reading (continued)
842–846		289–303	2.6	2.4–3.4	
847–849		304–317	2.7	2.4–3.4	
850–853		318–330	2.8	2.5–3.5	
854–856		331–343	2.9	2.5–3.5	
857–858		344–355	3.0	2.6–3.6	
859–861		356–367	3.1	2.6–3.7	
862–864		368–378	3.2	2.7–3.8	
865–865		379–389	3.3	2.7–3.8	
865–867		390–399	3.4	2.8–3.9	
868–868		400–409	3.5	2.8–4.0	
869–869		410–419	3.6	2.8–4.1	
870–870		420–428	3.7	2.9–4.2	
870–872		429–437	3.8	2.9–4.3	
873–873		438–446	3.9	3.0–4.4	
874–874		447–455	4.0	3.0–4.5	
876–876		456–465	4.1	3.0–4.6	
876–877		466–474	4.2	3.1–4.7	
877–878		475–483	4.3	3.1–4.8	
878–878		484–492	4.4	3.2–4.9	
878–879		493–502	4.5	3.2–5.0	
879–880		503–511	4.6	3.2–5.1	
880–881		512–521	4.7	3.3–5.2	
881–882		522–531	4.8	3.3–5.2	
882–882		532–542	4.9	3.4–5.3	
882–883		543–552	5.0	3.4–5.4	
883–884		553–563	5.1	3.5–5.5	

Table 62: Relating STAR Early Literacy Enterprise Scores to STAR Reading Scores (Continued)

STAR Early Literacy Enterprise		STAR Reading			Recommended Assessment(s)
Scale Score Range	**Literacy Classification**	**Scale Score Range**	**GE**	**ZPD Range**	
884–884	Probable Reader (continued)	564–574	5.2	3.5–5.5	STAR Reading (continued)
885–885		575–586	5.3	3.6–5.6	
885–886		587–597	5.4	3.6–5.6	
886–886		598–609	5.5	3.7–5.7	
886–887		610–621	5.6	3.8–5.8	
887–887		622–633	5.7	3.8–5.9	
887–888		634–645	5.8	3.9–5.9	
888–888		646–658	5.9	3.9–6.0	
889+		659+	6.0	4.0–6.1	

In addition to providing approximate GE scores and ZPDs, this table helps determine whether a student previously tested with STAR Early Literacy Enterprise could complete a STAR Reading test.

The 2001 STAR Early Literacy Validation Study data indicate that a student with a STAR Early Literacy Enterprise scaled score of 682 or higher is likely able to complete a STAR Reading test without getting frustrated. However, teachers should consider both the scaled score and their knowledge of the student's reading proficiency when selecting a reading assessment. Moreover, although a student may be capable of taking STAR Reading, teachers may want to continue using STAR Early Literacy Enterprise with the student to diagnose strengths and weaknesses in literacy skills and to plan reading instruction.

Diagnostic Assessment

One of the most powerful features of STAR Early Literacy Enterprise is its ability to function as a diagnostic assessment as described in No Child Left Behind. The program can

- identify a student's specific areas of strength and weakness
- determine any difficulties that a student may have in learning to read
- identify the potential cause of the difficulties
- help teachers determine appropriate reading intervention strategies

The Diagnostic–Student Report (also called the Student Diagnostic Report Skill Set Scores) is the ideal way to identify students' strengths and weaknesses. It displays Scaled Scores and reading development stage classifications for individual students and categorizes all Skill Set Scores into a table of strengths and weaknesses in each of the ten literacy sub-domains. By referring to this report, teachers can develop instructional strategies that capitalize on students' strengths and help students overcome their weaknesses.

A related report, the Score Distribution Report, is a classroom-level summary of students' skills. It shows the number of students in each of four score categories for the 41 Skill Set Scores provided by STAR Early Literacy Enterprise. The Score Distribution Report allows teachers to group students for instruction by the skills that need improvement so that instruction can be both efficient and effective.

Progress Monitoring

Research has shown that students whose progress is monitored regularly make greater gains than other students. Progress monitoring, sometimes called curriculum based assessment or curriculum based measurement, is an assessment conducted on a routine basis, weekly or monthly, that shows how well students are moving toward goals that have been established for them. Information gathered from progress monitoring is used to compare expected and actual rates of learning and to modify instruction as needed.

Beyond improving instruction, progress monitoring is an appropriate way to meet the adequate yearly progress (AYP) requirement of No Child Left Behind. By monitoring students' progress on a regular basis, teachers will be able to identify those who are on track to meet yearly goals and those who will require additional learning opportunities.

STAR Early Literacy Enterprise is an ideal way to conduct progress monitoring assessment. It reports student level information at the skill, sub-domain, or scaled score level and can be used on a weekly basis. The item pool is large enough to support frequent assessment using alternate forms that reflect the current level of a student's literacy status. Students can use the program independently or with minimum supervision, and because STAR Early Literacy Enterprise is a computer-adaptive assessment, most tests will take 10 minutes or less, not including pre-test instructions and mouse training. Including the instructions and training, which are optional after the first test administration, a wide majority of students will finish in 15 minutes or less.

One of the keys to successful progress monitoring is establishing intermediate goals. With STAR Early Literacy Enterprise, this process is easily accomplished using the tables in the section of this manual entitled "Score Definitions" on page 110. In this section, teachers can see the scores that correspond to various percentiles. Scaled Scores can be used to establish broad literacy goals and determine if students are making adequate progress. If students are not achieving the intermediate goals, Sub-domain and Skill Set Scores can provide specific information that will allow the teacher to modify instruction in the areas in which students are not making adequate progress.

Beginning with the STAR Early Literacy version 3.3, Renaissance Place editions include graphical Annual Progress Reports. The report displays either individual or class scores from all tests administered within the current school year. Using this report, teachers can view students' progress in terms of absolute growth through comparison to Literacy Classifications or relative growth through comparison to benchmarks.

Users of prior versions of STAR Early Literacy can also use the Growth Report to facilitate progress monitoring. This class-level report lists every student's Sub-domain Scores and Scaled Scores in chronological order over two test administrations. The Growth Report also provides class averages of scores, score changes for the initial and most recent administrations, and students' Estimated Oral Reading Fluency (Est. ORF) Scores.

By administering STAR Early Literacy Enterprise repeatedly during the school year, changes in scores from one administration to another can be used to monitor the progress of individual students and of the class as a whole. The Growth Report is a class-level report that lists every student's Sub-domain Scores and Scaled Scores in chronological order over two test administrations.

Goal Setting for Student Progress Monitoring

By using STAR Early Literacy Enterprise on a regular basis, such as at the beginning of each month of the school year, teachers can monitor students' progress and make appropriate adjustments to instructional practices. Progress monitoring is an approach that has strong research support and has proven successful in a variety of educational settings.

STAR Early Literacy Enterprise is appropriate for progress monitoring because it typically takes 10 minutes or less to administer, it can be administered frequently, the results of the assessment can be graphed to show growth, and the assessment comprises educationally relevant skills. The guidelines in this section will help teachers establish appropriate end-of-year goals and the intermediate district growth expectations needed to achieve these goals.

The Scaled Score is the most appropriate measurement to use for progress monitoring. This score ranges from 300 to 900 and represents an overview of a student's literacy status because it comprises the results of adaptive assessment in the sub-domains covered by STAR Early Literacy Enterprise.

Periodic Improvement

Data from the STAR Early Literacy Enterprise Validation Study make it possible to estimate the progress students should make on a monthly and annual basis. It is important to keep in mind that changes in Scaled Scores vary from grade-to-grade and from one achievement level to another. Generally speaking, younger students will probably make larger gains than older students because their literacy status is changing so quickly. The same is true of students who are at lower achievement levels when compared to those who are at higher achievement levels.

These differential changes are not unique to STAR Early Literacy Enterprise, but are typical of virtually all assessments, both formal and informal.

Table 63 shows the monthly gains that will move students from their current percentile to a somewhat higher percentile in the subsequent grade. For example, students in the 25th percentile in first grade will end up in a higher percentile in the second grade if their monthly score on STAR Early Literacy Enterprise increases on average by 1.5 percent. The percentages in the table are estimates, and meaningful variations can exist among students.

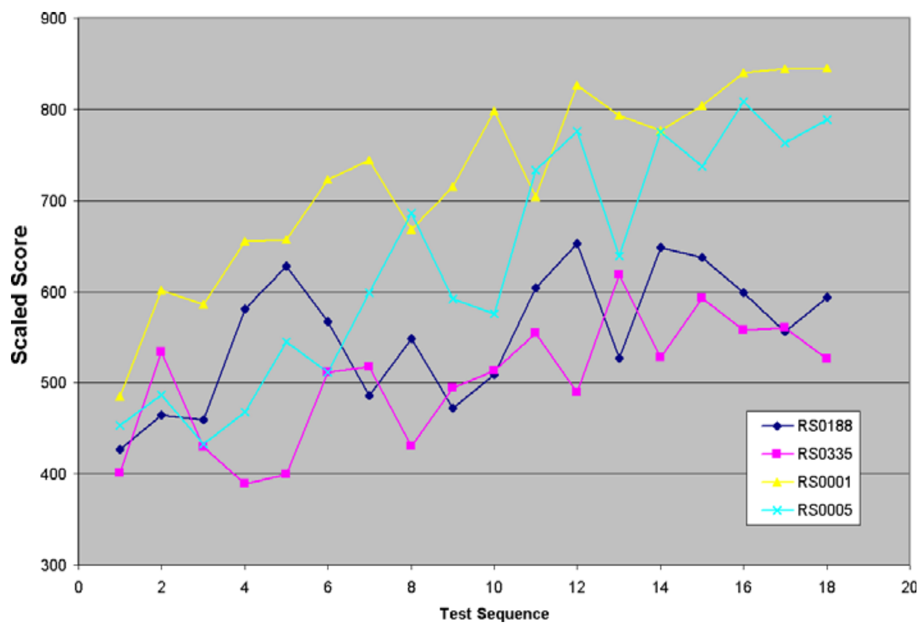
Table 63: Monthly Scaled Score Percentage Increases by Grade

Grade	K	1	2	3
Lower-achieving students (25th percentile)	4.5	1.5	1.0	1.0
Average-achieving students (50th percentile)	4.0	1.0	0.5	0.4

Students in the higher percentiles will make relatively small percentage gains, if any at all, because of a ceiling effect. The ceiling effect is a characteristic of most assessments because there is a limit to the upper range of performance for many tasks. A good example of this is oral fluency. The speed at which text can be read aloud is limited by the mechanical aspects of speaking.

The month to month Scaled Scores for a student are unlikely to move upward consistently. Figure 15 shows the score trajectories for four different students for eighteen administrations of STAR Early Literacy. All of the students showed gains from initial to final assessments, but the trajectory of growth was erratic. This growth pattern is to be expected and reflects the measurement error in tests and the fluctuation in students' test performance from one occasion to another. A decline in Scaled Score from one test to the next is not a matter of concern unless it is larger than two standard errors of measurement. Intermittent score declines and erratic trajectories are not unique to STAR Early Literacy. They happen with all other tests that are administered at frequent intervals. A good example of this is the progress graph reported in "Developments in Curriculum-Based Measurement" (Deno, S. *Journal of Special Education*, 37, 184–192).

Figure 15: Four Trajectories of STAR Early Literacy Scaled Scores



Adequate Yearly Progress

Establishing adequate yearly progress goals for students can also be accomplished using data from STAR Early Literacy Enterprise. The process requires several steps, but it is relatively straightforward.

- **Establish a baseline.** Use a recent STAR Early Literacy Enterprise Scaled Score as the baseline. A score from a single assessment may be used, but a more dependable baseline would be obtained by using the average of several recent administrations of STAR Early Literacy Enterprise or a Consolidated Scaled Score.

- **Set a long-term goal**. Choose one of the following Scaled Scores as the goal. The scores represent the 50th percentile of the students in grades K through 3 in the Validation Study. Even though STAR Early Literacy Enterprise is not a normed test, the Validation Study sample provides a fixed reference group that is useful as a percentile frame of reference. This percentile was chosen because it represents the lower range of scores for students who are considered Proficient readers in many states. The goal should have a realistic time frame, given the current literacy status of the student. For example, a student in the middle of first grade with a Scaled Score of 550 is unlikely to reach the 50th percentile by the end of the first grade. It is more likely that the student will reach the 50th percentile by the end of second grade or maybe even third grade. In this case, the most reasonable time frame is the end of third grade.

Grade	K	1	2	3
50th Percentile Score	585	763	816	841

- **Calculate the overall district growth expectation.** Subtract the current Scaled Score from the goal score. In the case of our example, the calculation would be 841 – 550 = 291. The student's score would have to increase by 291 points in order to reach the 50th percentile by the end of third grade.
- **Calculate the monthly district growth expectation**. Divide the score increase by the number of instructional months available before the goal time. In this case, the number of months would be 4 + 9 + 9 = 22 because four months would remain in the current year and nine months would remain in each of second and third grade. The monthly gain in Scaled Score would be approximately 15 points.

  Set year-end goals. Calculate the end-of-year scores needed to achieve the long-range goal. For our example, the year end goals are shown below.

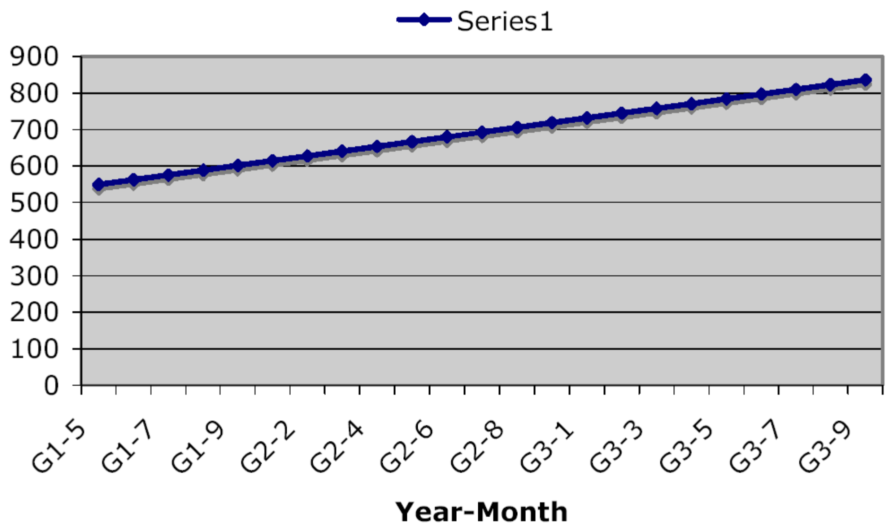
  End of Grade 1 (4 × 13) + 550 = 602

  End of Grade 2 (9 × 13) + 602 = 719

  End of Grade 3 (9 × 13) + 719 = 836

In addition to establishing end-of-year goal scores, it is useful to create a graph showing the baseline score and the intermediate end-of year-scores. Our sample data are shown on Figure 16.

Figure 16: Adequate Yearly Progress Goals



Students' actual monthly scores can be plotted against the anticipated growth trajectory. Although there will undoubtedly be significant month-to-month variation, as shown in Figure 15, the overall trend should conform to the expected growth line on the graph.

Outcome Measurement

Outcome measurement is closely associated with progress monitoring. It is the process of establishing achievable year-end goals that reflect a student's current literacy status and instructional needs. Movement toward these year-end goals is measured through progress monitoring assessment and culminates in an end of year assessment.

The tables in the section of this manual entitled "Score Definitions" on page 110, in combination with students' scores on STAR Early Literacy Enterprise, are extremely useful in establishing year-end goals for outcome measurement.

Consider the example of Juana, a second grade student. The results of the first administration of STAR Early Literacy Enterprise in September showed that her Scaled Score was 785. This puts her somewhere between the 25th and 50th percentile, as shown by Table 52 on page 108. This table shows Scaled Scores by grade and percentile. Juana's teacher, in collaboration with Juana's parents, has chosen 830 as the target Scaled Score for the end of the year.

This score corresponds to a percentile between 50 and 75, a goal that is both ambitious and achievable. Moreover, if Juana achieves this goal, she will be reading at grade level and will be prepared for the upcoming challenge of third grade. Juana's progress toward this goal can be easily monitored with STAR Early Literacy Enterprise.

STAR Early Literacy Enterprise and Instructional Planning

Assessment serves a number of purposes, but none is so important as informing instruction. STAR Early Literacy Enterprise provides skill-level data that allows teachers to develop comprehensive instructional plans for individual students. No other assessment gathers so much information in such a brief period of time and presents it in such an understandable way.

The Diagnostic–Student Report (also called the Student Diagnostic Report Skill Set Scores) is the ideal source of information for instructional planning. This individual student report lists each skill in STAR Early Literacy Enterprise, categorized by sub-domain, and indicates the relative strength or weakness of each skill. In addition, the Diagnostic–Student Report displays the Scaled Score, reading development stage, and Estimated Oral Reading Fluency Score for the student.

On the Diagnostic–Student Report, skills are placed in columns corresponding to the student's score on each skill. The cells represent the following score ranges: < 25%, 25–49%, 50–75%, and > 75%. If a skill is in the > 75% column, students are approaching mastery. Conversely, skills in the < 25% column have not yet begun to emerge.

Teachers can follow several steps in order to take the greatest advantage of the information in the Diagnostic–Student Report. The first is to skim the report to gain an overall idea of a student's general reading ability (Scaled Score) and relative strengths and weaknesses at the skill level.

This overall review will help the teacher decide which students are most in need of special instruction and which students seem to be benefiting from regular instruction. Teachers can base this decision on the average Scale Score for the class using the Summary Report or the information contained in Table 52 on page 108. The Summary Report shows individual Scaled Scores, Sub-domain Scores, Estimated Oral Reading Fluency Scores, and reading development stage classifications as well as class averages. It is a good source of information for grouping students for specific skill instruction.

Next, the teacher must decide how to prioritize instructional interventions. This decision is not so obvious as it might seem at first. For example, the students who have the lowest Scaled Scores might not be those who should receive targeted instruction first. Because their needs are so great, the teacher might want to plan a long-term intervention in order to provide these students with the intensive support they need. In contrast, a student whose Scaled Score is somewhat below what is expected, might be a better candidate for immediate intervention. Providing this student with additional instruction in the sub-domain or skill areas that are lowest might remedy the problem quickly so the student can progress with the rest of the group.

That is not to suggest that other students might not qualify for immediate intervention. It is the teacher's role to make the appropriate decision based on the best information available. STAR Early Literacy Enterprise can provide that information.

In terms of intervention strategies for students who are below expectations for a number of skills, the teacher must make another critical decision, where to begin. Again, the lowest skill or sub-domain scores might not be the appropriate starting point. Instead, the teacher should decide which skills should be addressed first, based on the literacy status and needs of a student as well as the importance of a skill.

Conventional reading and writing skills that are developed in the years from birth to age 5 have a consistently strong relationship with later conventional literacy skills. The National Early Literacy Panel's Developing Early Literacy Report (published by the National Institute for Literacy) identifies early literacy variables and skills that have medium to large predictive relationships with later measures of literacy development.

STAR Early Literacy Enterprise addressees the following predictive early literacy variables and skills:

<table>
<tr><td>Alphabet knowledge</td><td>Knowledge of the names and sounds associated with printed letters.</td><td rowspan="3">Medium to large predictive relationships with later measures of literacy development; variables maintain their predictive power even when other variables, such as IQ or SES, were accounted for.</td></tr>
<tr><td>Phonological awareness</td><td>The ability to detect, manipulate, or analyze the auditory aspects of spoken language (including the ability to distinguish or segment words, syllables, or phonemes), independent of meaning.</td></tr>
<tr><td>Phonological memory</td><td>The ability to remember spoken information for a short period of time.</td></tr>
<tr><td>Visual processing</td><td>The ability to match or discriminate visually presented symbols.</td><td>Moderately correlated with at least one measure of later literacy achievement but either did not maintain this predictive power when other important contextual variables were accounted for or have not yet been evaluated by researchers in this way.</td></tr>
</table>

In addition, STAR Early Literacy Enterprise Teacher Activities address the following predictive early literacy variables and skills:

<table>
<tr><td>Rapid automatic naming of letters or digits</td><td>The ability to rapidly name a sequence of random letters or digits.</td><td rowspan="3">Medium to large predictive relationships with later measures of literacy development; variables maintain their predictive power even when other variables, such as IQ or SES, were accounted for.</td></tr>
<tr><td>Rapid automatic naming of objects or colors</td><td>The ability to rapidly name a sequence of repeating random sets of pictures of objects (e.g., car, tree, house, man) or colors.</td></tr>
<tr><td>Writing or writing name</td><td>The ability to write letters in isolation on request or to write one's own name.</td></tr>
<tr><td>Concepts about print</td><td>Knowledge of print conventions (e.g., left–right, front–back) and concepts (book cover, author, text).</td><td rowspan="2">Moderately correlated with at least one measure of later literacy achievement but either did not maintain this predictive power when other important contextual variables were accounted for or have not yet been evaluated by researchers in this way.</td></tr>
<tr><td>Oral language</td><td>The ability to produce or comprehend spoken language, including vocabulary and grammar.</td></tr>
</table>

The other skills featured in STAR Early Literacy Enterprise are certainly important, particularly those dealing with vocabulary and comprehension, but the skills listed above are the stepping stones in learning to read. If a student scores in the < 25% range on any of these skills, they should be the focus of intervention.

In addition to being skill-specific, the information provided by the Diagnostic–Student Report can be used with any reading curriculum. The skills in STAR Early Literacy Enterprise reflect the findings of the National Reading Panel, are based on the most current research on early reading, and they are consistent with both top-down and bottom-up approaches to reading instruction. As an example, no matter what instructional approach is used, students must be able to associate letters with their sounds in order to learn to decode words and eventually read them automatically.

Measuring Growth

In the primary classroom, students' literacy skills are changing at a faster rate than at any other time in school. These changes differ from student to student with respect to sequence as well as rate, and the development of skills is affected by both the home and school environment. Given the relationship of early literacy skills to later reading success, assessing the growth of students' early literacy skills is of critical importance. STAR Early Literacy Enterprise allows the teacher to assess growth in a number of ways that can inform instruction and evaluate the effectiveness of educational interventions.

Absolute Growth and Relative Growth

It is important to distinguish between two types of academic growth (or gains) that may be evidenced in test results. Absolute growth reflects any and all growth that has occurred. Relative growth reflects only growth that is above and beyond "normal" growth (i.e., beyond typical growth in a reference or norming group). In general, norm-referenced scores, such as percentiles, indicate relative growth. Scaled Scores, Sub-domain Scores, and Skill Set Scores all reflect absolute growth. The Growth Report in STAR Early Literacy Enterprise provides administrators, teachers, and parents with information about students' absolute and relative growth in literacy skills.

Information about students' absolute growth is more useful than relative growth because it helps educators and parents evaluate the effectiveness of learning activities and make the adjustments that will promote appropriate development.

The Pretest-Posttest Paradigm for Measuring Growth

For many classroom purposes, STAR Early Literacy Enterprise is the ideal tool to measure the progress of individual students. It is self-administered for most students, can be administered frequently, and provides a snapshot of a student's abilities over time.

On occasion, however, educators may want to use STAR Early Literacy Enterprise to evaluate the effectiveness of a specific intervention, such as a new textbook or instructional philosophy. In general, most educational program evaluation designs attempt to determine if relative growth has occurred. That is, they are attempting to measure the impact of the intervention, or program, above and beyond normal growth (i.e., above and beyond what you would expect to occur without the intervention). This approach is not easily applicable using STAR Early Literacy Enterprise because, by design, it does not provide an index of "normal" growth. STAR Early Literacy Enterprise can be used to evaluate growth, however, by means of a pretest-posttest paradigm with a control group.

The logical method for measuring growth (i.e., measuring effectiveness of educational interventions) is through the use of a pretest-posttest design. In such a design, each student is administered a test prior to the beginning of the intervention to establish a baseline measure.

Then, each student is measured again at a later point in time (usually with a different, but equated, "form" of the same test) to see if the intervention is providing the desired outcome. The follow-up measurement may be at the end of the intervention, or may be done periodically throughout the course of the new program. Certainly, all of the issues relating to the adequacy of the test itself (e.g., in terms of core issues of reliability and validity) are applicable in order for this type of research to work properly. One key factor in conducting pretest-posttest designs is that if the same test is used both times, then the results may be compromised due to students having previously been exposed to the test items. In an ideal situation, equivalent (parallel) tests with no items in common should be administered. As an alternative to parallel tests, subsequent administration of a computerized adaptive test, such as STAR Early Literacy Enterprise, is useful for these types of assessments since it ensures that students get psychometrically comparable scores on pretest and posttest administrations, with few or no common items.

It is important to note that, for evaluation purposes, growth is best measured at a group level, such as a classroom or grade level. This is because at the individual student level, there are technical issues of unreliability associated with growth (gain) scores.

Pretest-Posttest with Control Group Design

In the "classic" implementation of a pretest-posttest design, the group (classroom or school) receiving the new intervention is referred to as the experimental group. A second matched group that does not receive the intervention is referred to as the control group. The control group follows the same pretesting and posttesting pattern in order to serve as a baseline for "normal" growth (without the intervention). Growth is indicated when the difference between the groups' average (mean) scores (computed as posttest mean score minus pretest mean score) is positive. Because it is likely that growth will occur even if the program (or intervention) is ineffective, the program's effectiveness (or impact) is measured when the growth for the experimental group is significantly greater than the growth for the control group.

Using Scores to Measure Growth

Three basic pieces of score information are available from STAR Early Literacy Enterprise: Scaled Scores, Sub-domain Scores, and Skill Set Scores. Each score reflects a different aspect of learning and provides a different perspective on students' skills. Thus, it is important to note the differences among the three types of scores and consider how each score can serve to measure growth.

Scaled Scores

The best estimate of a student's overall reading ability at a given time is the scaled score. These scores represent the student's reading ability on a continuous vertical scale that spans all grade levels (pre-kindergarten through 3).

The underlying vertical scale was derived as part of the test development process. In adaptive testing, students can receive different sets of items and still receive a comparable Scaled Score that represents their unique underlying ability level. Because Scaled Scores essentially map a student to a specific location on the underlying ability continuum, they can be useful in measuring absolute growth, and they are included in the Growth Report in STAR Early Literacy Enterprise.

Sub-domain Scores

Sub-domain Scores express a student's performance in terms of degrees of proficiency in each of the ten literacy sub-domains that comprise the STAR Early Literacy Enterprise assessment. Every item in the STAR Early Literacy Enterprise adaptive item bank comes from one, and only one, of the literacy sub-domains. The score for a specific sub-domain, such as Phonemic Awareness (PA), is a direct estimate of the percentage of STAR Early Literacy Enterprise Phonemic Awareness items the student could answer correctly if all of them were administered. Thus, a student with a Phonemic Awareness Sub-domain Score of 75 can be expected to be able to answer 75 percent of all the PA items correctly.

Sub-domain Scores are directly related to Scaled Scores because both types of scores are derived directly from the Rasch ability scale that is used internally in STAR Early Literacy Enterprise. Like Scaled Scores, Sub-domain Scores can be useful for measuring absolute growth, and changes in them are included in the Growth Report. The difference between a student's Sub-domain Scores on different occasions is itself a percentage. An increase of 10 points on a particular Sub-domain Score means that the student is estimated to be able to answer 10 percent more items correctly than was previously the case.

Skill Set Scores

Skill Set Scores are proficiency estimates like Sub-domain Scores, as described above, except for their frame of reference: while Sub-domain Scores estimate proficiency in an entire literacy sub-domain, Skill Set Scores estimate proficiency on the items of just one of the 41 skill sets (see page 16) that comprise the STAR Early Literacy Enterprise assessment. The Skill Set Score for a specific Skill, such as GR03 (differentiating letters), is a direct estimate of the percentage of items from that Skill the student could answer correctly if all of them were administered. Thus, the interpretation of Skill Set Scores is identical to that of Sub-domain Scores, except for the reference to a smaller set of literacy skills.

Skill Set Scores can be used for measuring absolute growth in the same manner described above for Sub-domain Scores. Although they are useful indicators of growth, changes in Skill Set Scores are not included in the Growth Report. The Diagnostic–Student Report (also called the Student Diagnostic Report Skill Set

Scores) contains a list of each Skill Set Score categorized in a way that makes it easy to identify students' strengths and weaknesses.

Estimated Oral Reading Fluency Scores

Estimated Oral Reading Fluency (Est. ORF) is an estimate of a student's ability to read words quickly and accurately in order to comprehend text efficiently. Students with oral reading fluency demonstrate accurate decoding, automatic word recognition, and appropriate use of the rhythmic aspects of language (e.g., intonation, phrasing, pitch, and emphasis). Est. ORF is reported in correct words per minute, and is based on a known relationship between STAR Early Literacy Enterprise performance and oral reading fluency.

The Estimated Oral Reading Fluency Score is included on the Growth Report, Screening Report, Summary Report, and Diagnostic–Student Report (also called the Student Diagnostic Report Skill Set Scores). See Table 66 on page 161 for Est. ORF Scores for selected STAR Early Literacy Enterprise Scaled Score (SS) units.

Student Growth Percentile (SGP)

Because STAR Early Literacy is so widely used, Renaissance Learning has data for millions of testing events. With these scores, we are able to calculate growth norms. In other words, we can approximate how much growth is typical for students of different achievement levels in different grades from one time period to another. Renaissance Learning first incorporated growth modeling into STAR Early Literacy reporting in 2008 via decile-based growth norms. SGPs represent the latest advancement in helping educators understand student growth. SGPs are available in STAR Early Literacy Enterprise for grades K–3.

SGPs are a normative 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. SGPs from STAR Early Literacy provide a measure of how a student changed from one STAR testing window[19] to the next, relative to other students with similar starting STAR Early 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 to another exceeds the growth of 70% of students in the same grade with a similar beginning (pretest) STAR Early Literacy score.

In applying the SGP approach to STAR data, Renaissance Learning has worked closely with the lead developer of SGP, Dr. Damian Betebenner, of the Center for

19. We collect data for our growth norms during three different time periods: fall, winter, and spring. More information about these time periods is provided later in this section.

Assessment, as well as technical advisor Dr. Daniel Bolt, an expert in quantitative methods and educational measurement from the University of Wisconsin-Madison. Because SGP was initially developed for measuring growth on state tests across years, applying the SGP approach to interim, within-year assessment data involved a number of technical challenges, primarily the differences regarding how STAR Early Literacy and state tests are administered. State summative tests are typically administered once a year, at approximately the same time, to all students. On the other hand, STAR Early Literacy is much more flexible, and may be administered to students as often as weekly. Decisions on when to administer and which students will participate are left to local educators. Most commonly, schools use STAR Early Literacy as a screening and benchmarking test for all or nearly all students 2–4 times per year. Students requiring more frequent progress monitoring may take STAR Early Literacy on a more frequent basis to inform instructional decisions, such as whether the student is responding adequately to an intervention. Because of this flexibility, not all students necessarily take STAR Early Literacy at the same time; the number and dates of administration may vary from one student to the next. SGP is calculated for students who have taken at least two tests within different testing windows. It uses the most recent test and at least one prior test from an earlier testing window (Fall, Winter, or Spring). The calculation uses the first test in the Fall, the test closest to January 15 in Winter, and the last test in Spring. Only tests taken in the last 18 months are used in the calculation.

To calculate Student Growth Percentiles, Renaissance Learning collected hosted student data from the two most recent school years (2011–12 and 2012–13). Table 64 (on the next page) has details on demographics of these students.

Table 64: Sample Characteristics, STAR Early Literacy SGP Study

		Sample %		
		Fall to Spring (n = 697,084)	**Fall to Winter (n = 688,938)**	**Winter to Spring (n = 802,472)**
Geographic Region	Midwest	20.8%	20.7%	22.3%
	Northeast	8.5%	9.3%	9.0%
	South	54.7%	53.7%	53.0%
	West	16.0%	16.3%	15.7%
	Response Rate	98.8%	98.7%	98.7%
School Type	Public	97.7%	97.6%	97.7%
	Private, Catholic	1.5%	1.6%	1.6%
	Private, Other	0.8%	0.8%	0.7%
	Response Rate	94.3%	94.1%	93.9%
School Enrollment	< 200	3.5%	3.6%	3.6%
	200–499	42.8%	43.2%	43.1%
	500–2,499	53.7%	53.2%	53.4%
	2,500 or more	0.0%	0.0%	0.0%
	Response Rate	96.3%	96.3%	95.8%
School Location	Urban	26.8%	25.9%	26.0%
	Suburban	25.1%	25.8%	26.4%
	Town	16.9%	17.1%	17.0%
	Rural	31.3%	31.1%	30.6%
	Response Rate	90.4%	90.6%	90.0%
Ethnic Group	Asian	2.6%	2.7%	2.7%
	Black	23.7%	22.8%	23.5%
	Hispanic	22.4%	21.8%	22.4%
	Native American	1.6%	1.8%	1.6%
	White	49.6%	50.9%	49.8%
	Response Rate	43.8%	42.7%	43.2%
Gender	Female	48.2%	48.0%	48.1%
	Male	51.8%	52.0%	51.9%
	Response Rate	82.0%	81.9%	81.7%

To address the variability in the number of days between students' pre and posttest dates, time had to be incorporated into our model. Taking this approach varies from the typical SGP approach in that it uses a combination of a student's pretest score along with his weekly rate of growth, instead of simply pre and posttest scaled scores. Quantile regression was applied to characterize the bivariate distribution of students' initial scores and weekly rates of growth. Students were grouped by grade and subject, and then quantile regression was used to associate every possible initial score and weekly growth rate combination with a percentile corresponding to the conditional distribution of weekly growth rate given the initial score. The result of these analyses was the creation of a look-up table in which initial STAR scores along with weekly growth rates are used as input to define a Student Growth Percentile for each grade, subject, and time period (e.g., fall to winter, winter to spring, fall to spring). The use of quantile regression techniques makes construction of such tables possible even though not all possible initial and ending score combinations were observed in the student data. In general, the quantile regression approach can be viewed as a type of smoothing in which information from neighboring score values (initial scores and weekly rates of growth) can be used to inform percentiles for hypothetical score combinations not yet observed. As such, application of the methodology allows us to look up any score combination to obtain the percentile cut points for the weekly growth rate conditional achievement distribution associated with the given initial score. These cut points are the percentiles of the conditional distribution associated with the student's prior achievement. Specifically, using the quantile regression results of the sixth-grade STAR Early Literacy weekly growth rate on fall scores, we can calculate estimates for the 1st, 2nd, 3rd,…99th percentiles of growth from fall to spring can be calculated. Using each of these cut points, we are able to calculate a Student Growth Percentile for every subject, grade, and score combination.

Choosing a Score to Measure Growth

The choice of Scaled Scores, Sub-domain Scores, or Skill Set Scores as a growth measure should reflect the purpose for which the scores will be used. At the classroom level, the teacher is most likely to be concerned about each type of score for different reasons.

- To gain an understanding of the general literacy level of a student, the Scaled Score is most important. The Scaled Score might also be useful if a teacher wants to evaluate the effectiveness of an instructional intervention with the group as a whole.
- To learn how individual students are doing with respect to a cluster of skills, the Sub-domain Score is most important. For example, a teacher might consider Sub-domain Scores in order to identify students who need additional opportunities to improve their vocabulary skills. These students might be grouped for an intervention that focuses on vocabulary. Sub-domain Scores could then be used to evaluate the effectiveness of the intervention.

- To learn how well students are doing on a specific skill, the Skill Set Score is most informative. A kindergarten teacher, for example, might pay special attention to Skill Set Scores in order to track the phonemic awareness of students whose Scaled Score is far below what is expected. The Scaled Score would serve as an overall indicator that the student is at risk for failing to learn to read, and because phonemic awareness is so closely related to reading success, the teacher would want to focus on this skill for additional instruction.

At the School or district level, the motivation in looking at growth is likely to be at the macro level—that is, the "big picture." Scaled Scores are probably the most appropriate measure of growth in such situations. Administrators might, for example, identify classrooms that have lower Scaled Scores than expected. In collaboration with the teacher, the administrator might be able to identify the source of the problem and provide the teacher with the resources needed to get students back on track.

No matter which scale is used to evaluate growth, if the evaluation is done at the level of the individual student, it will be important to keep two things in mind. The first is that frequent assessment using STAR Early Literacy Enterprise will provide a more dependable picture of a student's current status and progress. Typically, students do not progress in a continuously upward trajectory. Instead, they may have growth spurts or periods when scores actually decline. This is a reflection of both the typical developmental pattern of young students and measurement error.

The second thing to keep in mind is that individual score changes are much less reliable than average changes at the group level. Evaluation of individual changes should always include a consideration of the standard error of measurement. This is because on any educational test the student's obtained score is an estimate of his or her actual ability. The difference between true and estimated ability is a measurement error. Measurement errors are assumed to be normally distributed, with an average of zero, and a standard deviation equal to the standard error of measurement. About 68 percent of the time, the obtained score should lie within 1 standard error of measurement of the true score; about 95 percent of the time, it should lie within 2 standard errors of measurement of the true score.

For example, suppose a student's Scaled Score was 675,with a standard error of measurement of 25. Adding and subtracting 25 to the Scaled Score yields a range of scores from 650 to 700; this is a 68 percent confidence interval for the student's true score. STAR Early Literacy Enterprise Growth Reports include a graphical depiction of confidence intervals around Scaled Scores, by means of a horizontal line called an "error bar" extending to the left and right of each obtained Scaled Score. If the error bars from two different test administrations are overlapping, the score differences should not be viewed as real changes. To determine confidence intervals for Sub-domain Scores and Skill Set Scores, the reader should refer to the standard error of measurement tables in “Reliability and Measurement Precision” on page 44.

STAR Early Literacy Enterprise and the Reading First Initiative

The Reading First initiative is designed to help all students become successful and proficient readers. It requires districts to use materials that incorporate scientifically based reading research. Specifically, Reading First supports the five key components of literacy development outlined by the National Reading Panel: Phonemic Awareness, Phonics, Fluency, Vocabulary, and Comprehension. According to Reading First, effective reading interventions must support the core reading program with systematic and explicit instruction in one or more of these five research-supported areas. Reading First also targets interventions at the classroom instructional level, where students spend the majority of their time.

Every state may apply for and receive Reading First funds. These monies can span over a six-year period and are awarded competitively to local education agencies. In addition, targeted assistance grants will be awarded to states and local educational agencies based on evidence of significant increases in the percentage of grade 3 students reading at the proficient level and of improved reading skills for students in grades 1–3. Reading First applications must include plans to train teachers in the five essential components of reading, and to select and administer screening, diagnostic, and classroom-based instructional reading assessments to identify those children who may be at risk for reading failure. Contact the federal Department of Education (www.ed.gov) or your state Department of Education for more information about the Reading First initiative. Contact the Renaissance Learning Funding Center (www.renaissance.com) for assistance with applying for a Reading First Grant or to download a flowchart demonstrating how Renaissance Learning products align with Reading First.

STAR Early Literacy Enterprise aligns very well with the Reading First initiative. This Technical Manual demonstrates the ability of STAR Early Literacy Enterprise to assess four of the five key components of literacy development: Phonemic Awareness, Phonics, Vocabulary, and Comprehension. Hence, with the use of STAR Early Literacy Enterprise, educators can monitor students’ development in nearly all of the early literacy areas supported by the National Reading Panel and the Reading First initiative.

The Reading First Initiative also emphasizes the importance of documenting gains in reading achievement. States receiving Reading First funds must annually submit a report that, among other things, identifies schools and local educational agencies displaying the largest gains in reading achievement.

Using the pretest-posttest procedures described in this Technical Manual, educators can use STAR Early Literacy Enterprise Scaled Scores and Sub-domain Scores to demonstrate the effectiveness of their reading interventions and measure growth in literacy skills over time. The STAR Early Literacy Enterprise Growth Report may provide educators with a simple and accurate method for monitoring and documenting gains in literacy skills in accordance with Reading First.

Score Interpretation

This section discusses the interpretation of STAR Early Literacy Enterprise Scaled Scores, Sub-domain Scores, and Skill Set Scores. It is important to note that each of these three score types is derived directly from the student's estimated Rasch ability—a summary measure of a student's ability in the universe of literacy skills that STAR Early Literacy encompasses. Rasch ability is expressed on a real number scale. Typical values of Rasch ability in STAR Early Literacy range between –5.0 and +5.0. Because of their unfamiliar scale, Rasch ability scores do not appear on STAR Early Literacy score reports. Instead, scores in more familiar metrics are reported: scaled scores, sub-domain scores, and skill set scores.

Scaled Scores are direct but nonlinear conversions of Rasch ability to integer scores ranging between 300 and 900. This conversion to an integer scale was done to simplify the score scale used in the score reports. Scaled Scores below 310 and above 890 occur only rarely.

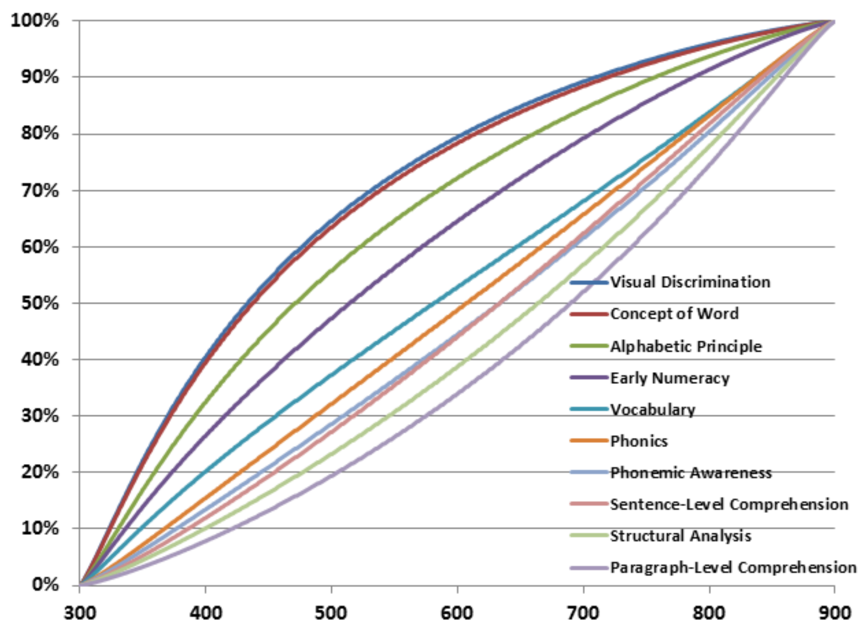
Sub-domain Scores are derived by applying the Rasch measurement model to estimate the percentage of items in each STAR Early Literacy Enterprise sub-domain that a student can answer correctly, given the value of his or her Rasch ability. Each of the ten Sub-domain Scores—AP, CW, EN, PC, PA, PH, SC, SA, VS and VO—is based on a unique subset of the items comprising the STAR Early Literacy Enterprise item bank. Because they are percentages, Sub-domain Scores range from 0 to 100.

There is a functional relationship between Sub-domain Scores and Scaled Scores. Within each of the 10 literacy sub-domains, for each possible value of the Scaled Score, there is a corresponding Sub-domain Score value.

The relationship of Sub-domain Scores to Scaled Scores is different, however, for each literacy Sub-domain. This is because the number of items, the average item difficulty, and the distribution of item difficulty differ from one Sub-domain to another.

Figure 17 illustrates the relationship of each of the 10 Sub-Domain Scores to Scaled Scores.[20] As the figure illustrates, Sub-domain Scores are higher for the Visual Discrimination (VS) Sub-domain than for any other Sub-domain. The second-highest Sub-domain Scores occur on Concept of Word (CW), while the lowest Sub-domain Scores occur for the Paragraph-Level Comprehension (PC) Sub-domain. This reinforces an important point: because of differences among the Sub-domains in terms of item difficulty, Sub-domain Scores for Visual Discrimination and Concept of Word will always be higher than the other Sub-Domain Scores; similarly, Paragraph-Level Comprehension Sub-domain Scores will always be the lowest ones. Differences among Sub-domain Scores for the other four Sub-domains are smaller by comparison and less consistent.

Figure 17: The Relationship of Sub-Domain Scores to Scaled Scores

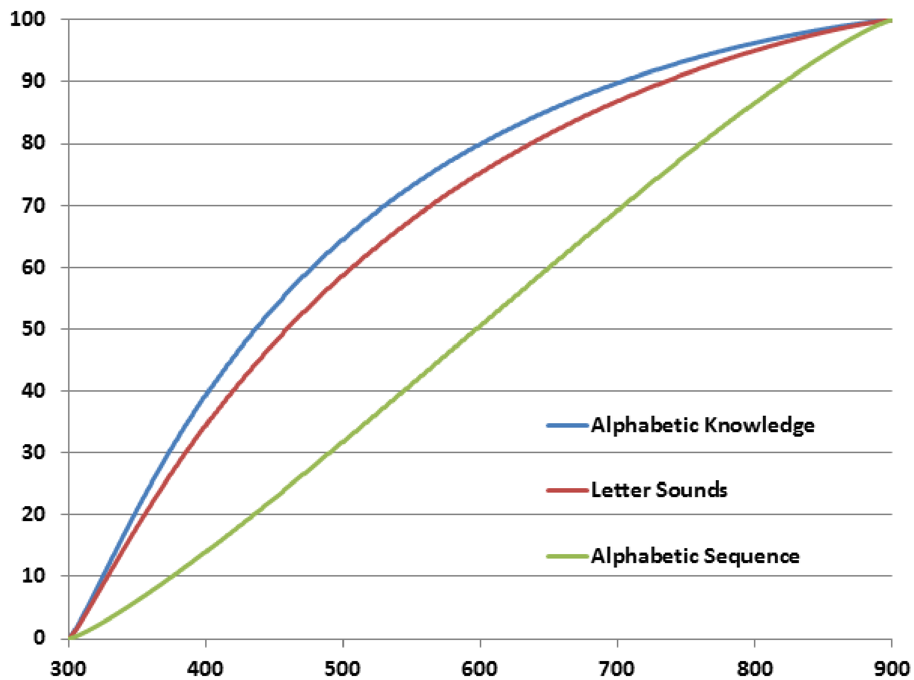


Skill Scores are derived in exactly the same way as Sub-domain Scores—by applying the Rasch measurement model to estimate the percentage of items in each STAR Early Literacy skill set that a student can answer correctly, given the value of his or her Rasch ability. Each Skill Set Score is based on a unique subset of the items in the item bank that make up the relevant literacy Sub-domain and Skill Set. As with Sub-domain Scores, there is a one-to-one relationship between Skill Set Scores and Scaled Scores. The relationship between Skill Set Scores and Scaled Scores is different for each skill because the number of items and the level and distribution of item difficulty differ from one skill to another.

20. Figures 17 through 26 display Sub-domain Scores and Skill Set Scores based on the initial STAR Early Literacy Enterprise item bank. Minor subsequent changes to the item bank will not appreciably affect the accuracy of the figures, or their interpretation.

Figure 18 shows the relationship of the three Alphabetic Principle Skill Set Scores to Scaled Scores. Each curve represents one of the three Alphabetic Principle Skill Set Scores. As the figure shows, Alphabetic Knowledge is the easiest of these skills, while Letter Sounds is somewhat more difficult and Alphabetic Sequence is the most difficult. Differences in the relative height of these three Skill Set Scores curves illustrate that there are material differences among the Alphabetic Principle skills in terms of difficulty and in terms of the Scaled Score level at which skill mastery can be expected.

Figure 18: Relationship of Alphabetic Principle Skill Set Scores to Scaled Scores



Because of differences among the skill sets in terms of item difficulty, there is a considerable degree of variation in the relationship of Skill Set Scores to Scaled Scores.

Figures 19 through 26 show the Skill Set Score-to-Scaled Score relationships for the Concept of Word, Early Numeracy, Paragraph-Level Comprehension, Phonemic Awareness, Phonics, Sentence-Level Comprehension, Structural Analysis, Visual Discrimination, and Vocabulary Skill Sets, respectively. Considered as a group, these figures demonstrate a substantial amount of differentiation among the Sub-domains and Skill Sets in terms of the relationship between Scaled Scores and Skill Set Scores.

Figure 19: Relationship of Concept of Word Skill Set Scores to Scaled Scores

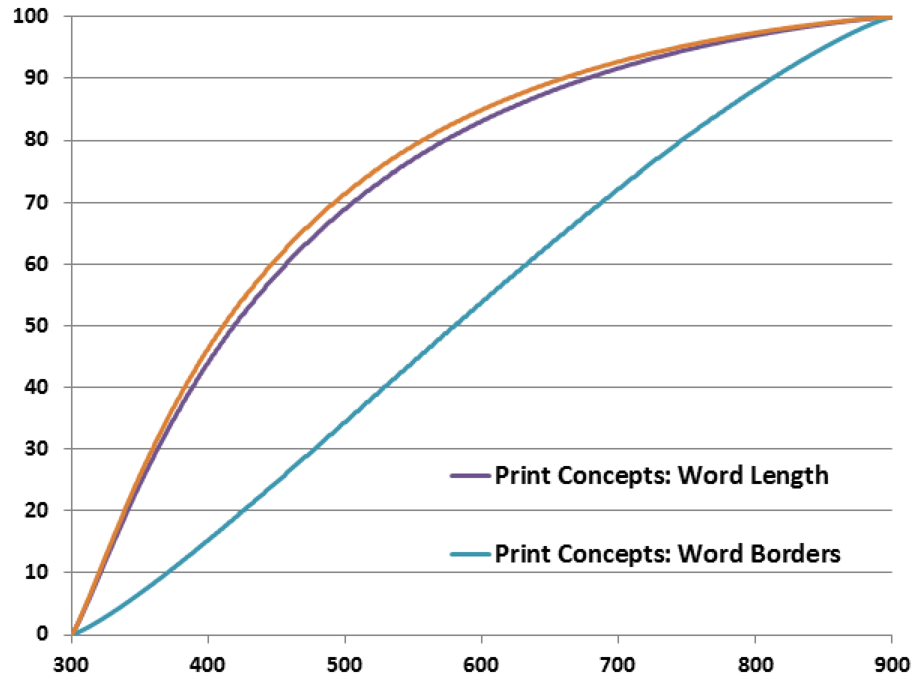


Figure 20: Relationship of Early Numeracy Skill Set Scores to Scaled Scores

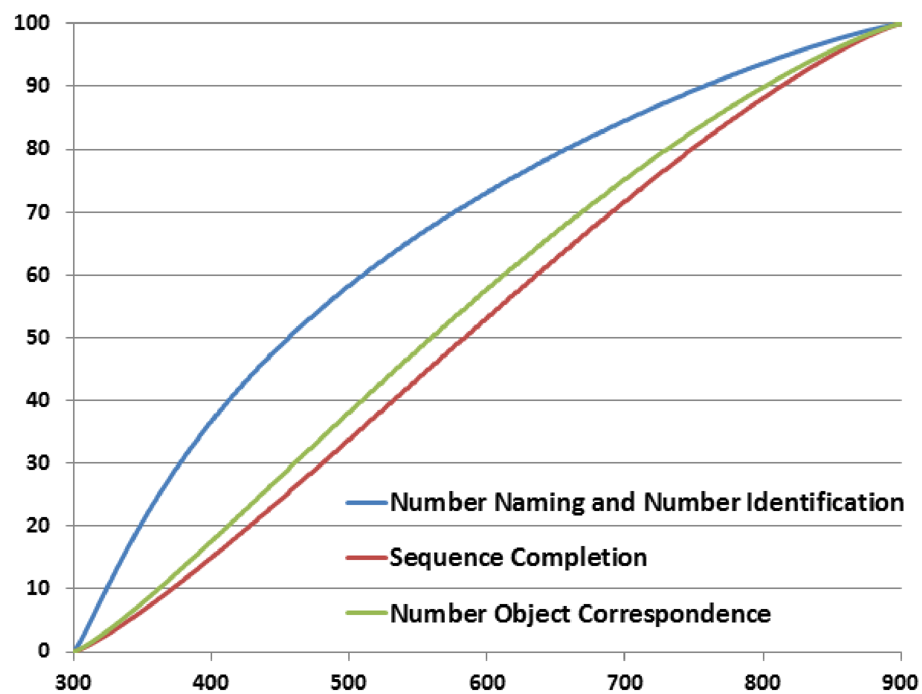


Figure 21 combines two related Sub-domains: Sentence-Level Comprehension and Paragraph-Level Comprehension. Each of these two Sub-domains contains only one skill set. As the figure shows, comprehension at the sentence level is the easier of the two skill sets, and comprehension of paragraphs is somewhat more difficult throughout most of the Scaled Score range.

Figure 21: Relationship of Sentence-Level and Paragraph-Level Comprehension Skill Set Scores to Scaled Scores

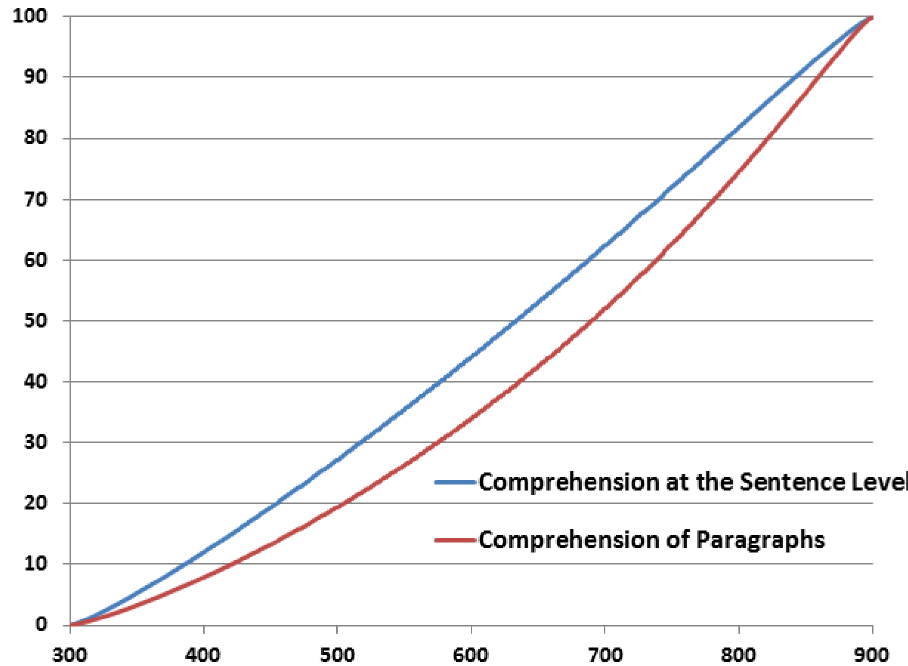


Figure 22: Relationship of Phonemic Awareness Skill Set Scores to Scaled Scores

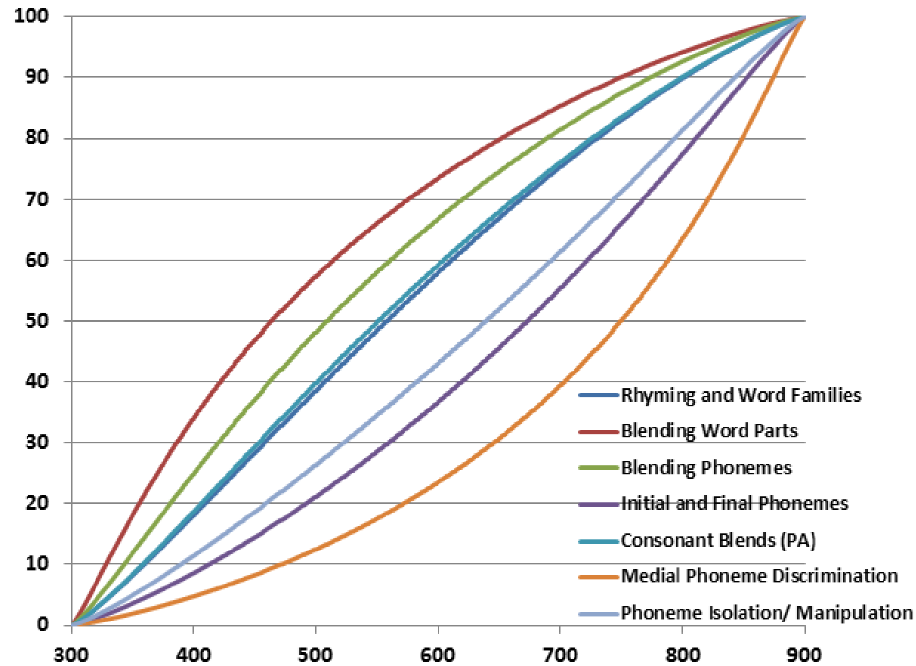


Figure 23: Relationship of Phonics Skill Set Scores to Scaled Scores

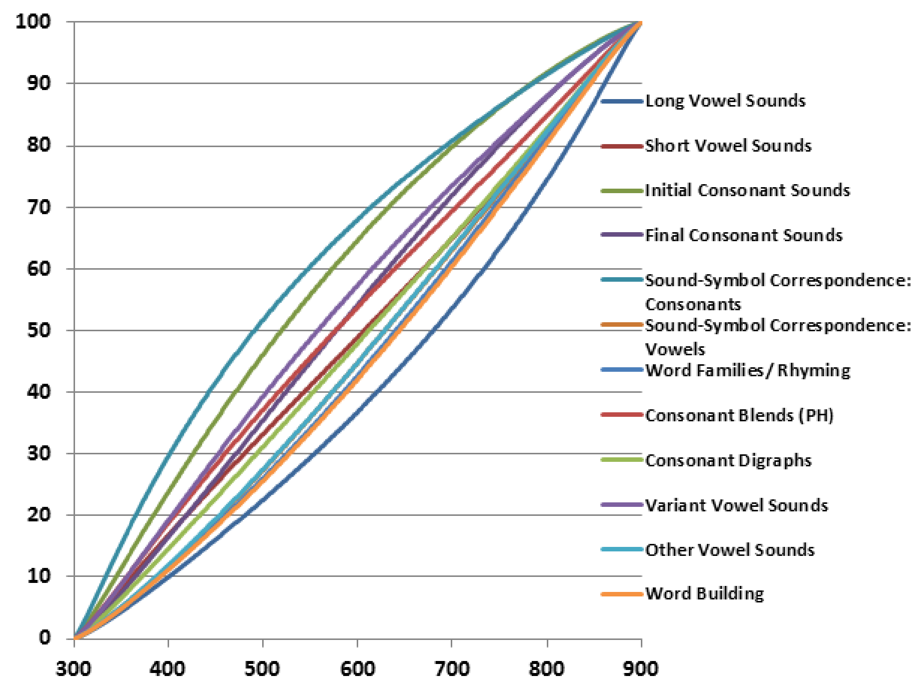


Figure 24: Relationship of Structural Analysis Skill Set Scores to Scaled Scores

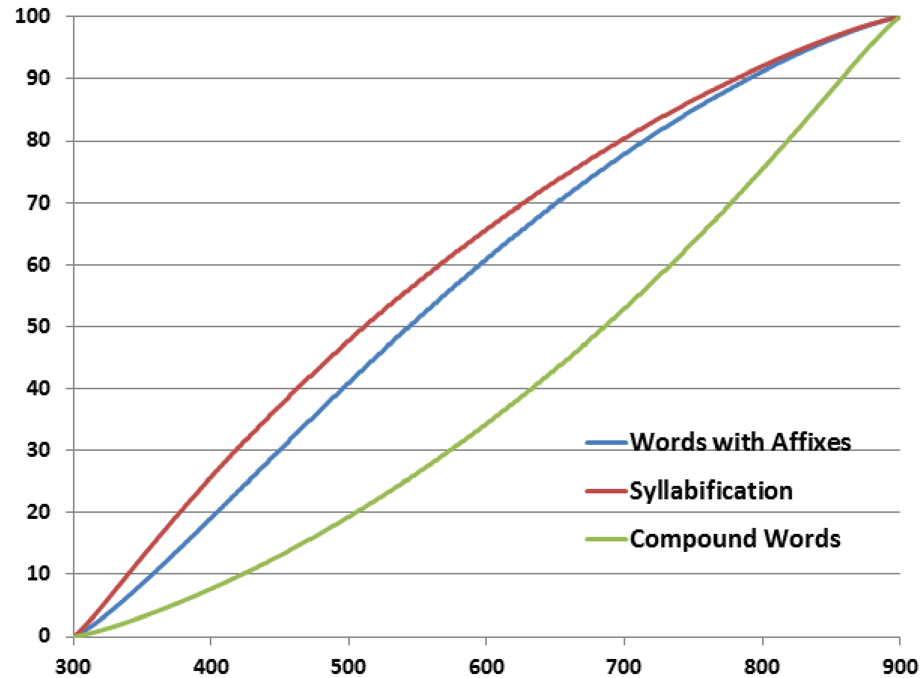


Figure 25: Relationship of Visual Discrimination Skill Set Scores to Scaled Scores

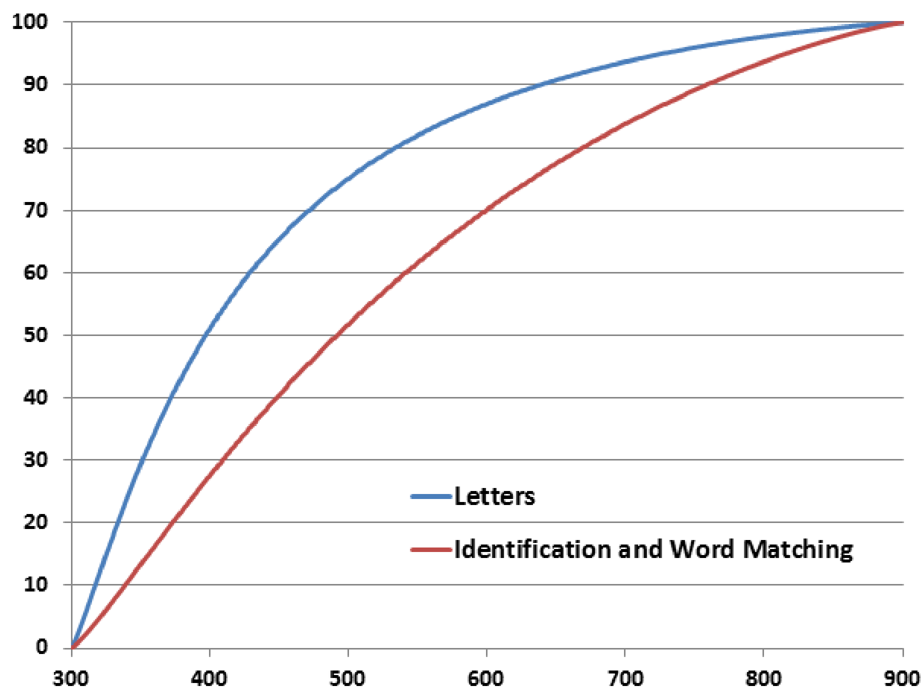


Figure 26: Relationship of Vocabulary Skill Set Scores to Scaled Scores

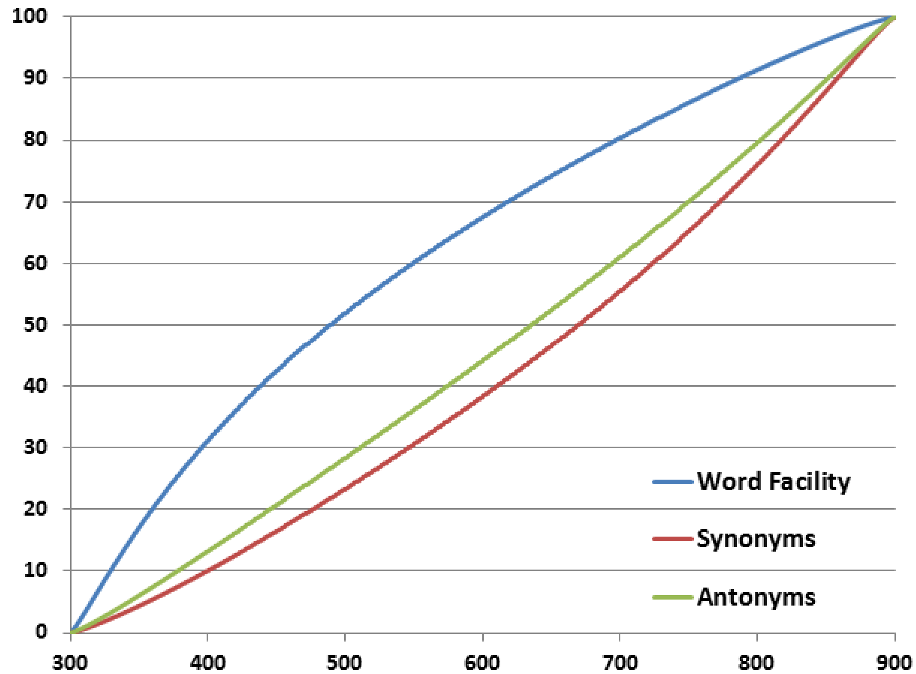


Table 65: STAR Early Literacy Enterprise Scaled Score to Percentile Rank Conversion by Grade (at Month 7 in the School Year)

PR	Grades			
	K	1	2	3
1	379	457	482	522
2	398	510	545	555
3	412	538	579	595
4	425	558	602	642
5	439	571	620	667
6	450	582	634	686
7	458	590	647	700
8	466	600	659	718
9	474	608	671	734
10	480	616	680	743
11	485	623	687	747
12	490	628	693	752
13	495	634	698	757
14	501	639	703	762
15	505	645	709	767
16	509	650	714	770
17	512	655	719	772
18	515	660	724	776
19	519	665	730	779
20	522	670	734	783
21	526	675	739	786
22	529	679	742	789
23	534	682	745	791
24	537	685	748	793

Table 65: STAR Early Literacy Enterprise Scaled Score to Percentile Rank Conversion by Grade (at Month 7 in the School Year) (Continued)

PR	Grades			
	K	1	2	3
25	540	689	752	796
26	543	691	755	798
27	546	694	757	802
28	549	697	761	805
29	553	700	764	808
30	556	703	767	809
31	559	706	770	810
32	562	708	773	812
33	565	711	775	814
34	567	714	777	816
35	570	717	780	818
36	572	720	782	820
37	575	723	784	821
38	577	725	786	822
39	580	728	788	824
40	583	731	790	825
41	586	733	792	827
42	589	736	794	828
43	591	738	796	830
44	594	741	798	831
45	596	743	800	832
46	599	745	802	833
47	602	747	804	834
48	605	750	806	836
49	608	752	808	837
50	611	754	810	838
51	613	756	811	840

Table 65: STAR Early Literacy Enterprise Scaled Score to Percentile Rank Conversion by Grade (at Month 7 in the School Year) (Continued)

PR	Grades			
	K	1	2	3
52	616	759	813	841
53	618	761	815	842
54	620	763	817	843
55	622	766	818	844
56	625	768	820	845
57	627	770	821	847
58	630	772	823	848
59	633	774	824	849
60	635	776	826	850
61	637	778	828	851
62	640	780	829	852
63	643	782	830	853
64	645	784	832	854
65	648	786	833	856
66	651	787	835	857
67	654	789	836	858
68	657	791	838	859
69	660	793	839	860
70	663	795	840	861
71	667	796	842	862
72	670	799	843	862
73	673	801	845	863
74	676	803	846	864
75	679	805	848	864
76	682	807	849	865
77	685	809	850	867
78	688	811	851	869

Table 65: STAR Early Literacy Enterprise Scaled Score to Percentile Rank Conversion by Grade (at Month 7 in the School Year) (Continued)

PR	Grades			
	K	1	2	3
79	692	813	853	870
80	695	815	854	870
81	698	818	856	871
82	701	819	857	872
83	705	821	859	872
84	708	823	860	872
85	712	826	861	873
86	716	828	862	873
87	720	830	864	875
88	725	832	865	877
89	732	835	869	879
90	741	838	871	880
91	750	841	872	880
92	759	845	874	881
93	767	849	879	881
94	778	853	880	886
95	786	858	881	889
96	795	862	886	889
97	807	869	889	890
98	822	875	890	894
99	840	882	895	896

Table 66: Estimated Oral Reading Fluency (Est. ORF) Given in Words Correct per Minute (WCPM) by Grade for Selected STAR Early Literacy Enterprise Scale Score Units (SEL SS)

SEL SS	Grade		
	1	2	3
300–460	0	0	0
480	0	1	0
500	0	3	0
520	1	6	3
540	5	7	6
560	8	9	8
580	11	13	10
600	13	15	12
620	16	18	16
640	18	20	18
660	21	22	21
680	24	25	25
700	26	27	28
720	29	30	32
740	34	34	37
760	42	40	44
780	51	49	50
800	62	58	56
820	74	71	65
840	88	87	78
860	140	112	103
880	142	175	150
900	142	175	170

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*. Manuscript in preparation.

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.

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.

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

E

F

G

H

I

K

L

M

N

S

About Renaissance Learning

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

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

Renaissance Learning™
P.O. Box 8036 | Wisconsin Rapids, WI 54495-8036
(800) 338-4204 | www.renaissance.com

L0796
43276.150804