RENAISSANCE LEARNING™

STAR Reading™ Technical Manual

STAR™ Reading

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

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

Copyright Notice

3/2015 SRRP

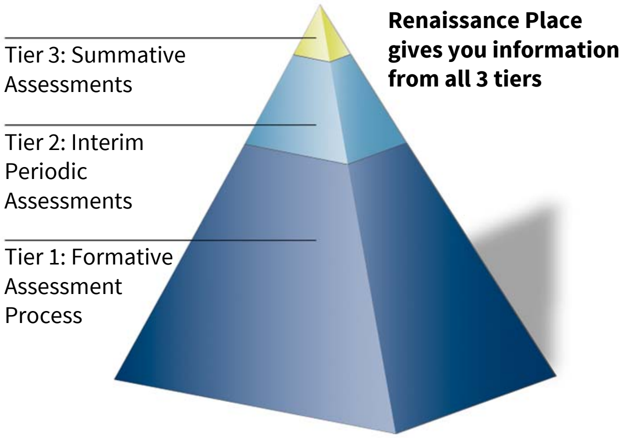
Contents

Introduction

STAR Reading: Screening and Progress-Monitoring Assessment

Beginning with the 2011–2012 school year, two different versions of STAR Reading are available for use in assessing the reading achievement of students in grades K–12. The Renaissance Place Edition of the STAR Reading computer-adaptive test and database allows teachers to assess students' reading comprehension and overall reading achievement in ten minutes or less. This computer-based progress-monitoring assessment provides immediate feedback to teachers and administrators on each student's reading development. The second version is the Renaissance Place Edition of STAR Reading Enterprise, which works in a similar fashion to STAR Reading, but measures and reports on a wide range of discrete reading skills that are aligned to state and national curriculum standards.

STAR Reading runs on the Renaissance Place platform, which stores three levels of critical student data: daily progress monitoring, periodic progress monitoring, and annual assessment results. Renaissance Learning identifies these three levels as Tier 1, Tier 2, and Tier 3, as described below.


Tier 3: Summative Assessments
Tier 2: Interim Periodic Assessments
Tier 1: Formative Assessment Process
Renaissance Place gives you information from all 3 tiers

Tier 1: Formative Assessment Process

A formative assessment process involves 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 laptops.

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, 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.

STAR Reading Purpose

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

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

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

- For students, STAR Reading software provides a challenging, interactive, and brief test that builds confidence in their reading ability.
- For teachers, the STAR Reading test facilitates individualized instruction by identifying children who need remediation or enrichment most.
- For principals, the STAR Reading 3 and higher Renaissance Place (RP) browser-based management program provides regular, accurate reports on performance at the class, grade, building, and district level.
- For district administrators and assessment specialists, the Renaissance Place program provides a wealth of reliable and timely data on reading growth at each school and districtwide. It also provides a valid basis for comparing data across schools, grades, and special student populations.

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

STAR Reading Enterprise

STAR Reading Enterprise is the same as STAR Reading, but with some enhanced features, including additional reports and expanded benchmark management.

In this manual, information that refers to Enterprise-only program functions will have the ENTERPRISE indicator next to them.

Design of STAR Reading

Three Generations of STAR Reading Assessments

The introduction of STAR Reading Enterprise in 2011 marks the third generation of STAR Reading assessments. The first generation consisted of STAR Reading version 1, which was a variable-length adaptive assessment of reading comprehension that employed a single item type: vocabulary-in-context items. STAR Reading's original item bank contained 838 such items. Although it was a breakthrough computer adaptive test, STAR Reading 1 was based on traditional test theory.

The second generation consisted of STAR Reading versions 2 through 4.4, including the current STAR Reading Service version. This second generation differed from the first in three major respects: It replaced traditional test theory with Item Response Theory (IRT) as the psychometric foundation for adaptive item selection and scoring; its test length was fixed at twenty-five items (rather than the variable length of version 1); and its content included a second item type: the original vocabulary in context items were augmented in grades 3–12 by the use of longer, authentic text passages for the last 5 items of each test. The second generation versions differed from one another primarily in terms of the size of their item banks, which grew from 1,409 items in version 2.0 to 2,048 items in version 4.4. Like the first generation of STAR Reading tests, the second generation continued to measure a single construct: reading comprehension.

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

Overarching Design Considerations

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

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

Another fundamental STAR Reading design decision involved the choice of the content and format of items for the test. Many types of stimulus and response procedures were explored, researched, discussed, and prototyped. These procedures included the traditional reading passage followed by sets of literal or inferential questions, previously published extended selections of text followed by open-ended questions requiring student-constructed answers, and several cloze-type procedures for passage presentation. While all of these procedures can be used to measure reading comprehension and overall reading achievement, the vocabulary-in-context format was finally selected as the primary item format. This decision was made for interrelated reasons of efficiency, breadth of construct coverage, and objectivity and simplicity of scoring. For students at grade levels K–2, the STAR Reading 3 and higher test administers 25 vocabulary-in-context items. For students at grade levels 3 and above, the test administers 20 vocabulary-in-context items in the first section of the test, and five authentic text passages with multiple-choice literal or inferential questions in the second section of the test.

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

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

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

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

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

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

Improvements to the STAR Reading Test in Version 2

Since the introduction of STAR Reading version 1 in 1996, STAR Reading has undergone a process of continuous research and improvement. Version 2 was an entirely new test, with new content and several technical innovations.

- The item bank was expanded from 838 test items distributed among 14 difficulty levels to 1,409 items graded into 54 difficulty levels.
- Test content was expanded as well. STAR Reading version 1 consisted of a single test section that measured reading comprehension through vocabulary-in-context questions. Versions 2 and higher add a section that uses authentic text passages to all tests administered to grades 3 and above to significantly enhance the test's ability to measure reading comprehension.
- The technical psychometric foundation for the test was improved. Versions 2 and higher are now based on Item Response Theory (IRT). The use of IRT permits more accurate calibration of item difficulty and more accurate measurement of students' reading ability.
- The Adaptive Branching process has likewise been improved. By using IRT, the STAR Reading 2 and higher tests effect an improvement in measurement efficiency.

- The length of the STAR Reading test has been shortened and standardized. Taking advantage of improved measurement efficiency, the STAR Reading 2 and higher tests administer just 25 questions to every student. At grade levels 3 and above, there are 20 vocabulary-in-context items and 5 authentic text passage items. At grade levels K–2, all 25 items are vocabulary-in-context items. In contrast, version 1 administered a variable number of items, ranging from 5–60. The average length of version 1 tests was 30 items per student.
- Like the STAR Reading 1 test before it, the STAR Reading 2 was nationally standardized prior to release. Therefore, its norm-referenced test scores represented the most recent benchmark available. Versions of STAR Reading between 2 and 4.3 all used the norms developed for version 2.

Improvements Specific to STAR Reading Versions 3 RP and Higher

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

This makes a number of new features possible:

- It makes it possible for multiple schools to share a central database, such as a district-level database. Records of students transferring between schools within the district will be maintained in the database; the only information that needs revision following a transfer is the student’s updated school and class assignments.
- The same database that contains STAR Reading data can contain data on other STAR tests, including STAR Early Literacy and STAR Math. The Renaissance Place program is a powerful information management program that allows you to manage all your district, school, personnel, parent, and student data in one place. Changes made to district, school, teacher, parent, and student data for any of these products, as well as other Renaissance Place software, are reflected in every other Renaissance Place program sharing the central database.
- Multiple levels of access are available, from the test administrator within a school or classroom to teachers, principals, district administrators, and even parents.
- Renaissance Place takes reporting to a new level. Not only can you generate reports from the student level all the way up to the school level, but you can also limit reports to specific groups, subgroups, and combinations of subgroups. This supports “disaggregated” reporting; for example, a report might be specific to students eligible for free or reduced lunch, to English language learners, or to students who fit both categories. It also supports

compiling reports by teacher, class, school, grade within a school, and many other criteria such as a specific date range. In addition, the Renaissance Place consolidated reports allow you to gather data from more than one program (such as STAR Reading and Accelerated Reader) at the teacher, class, school, and district level and display the information in one report.

- Since the Renaissance Place software is accessed through a web browser, teachers (and administrators) will be able to access the program from home—provided the district or school gives them that access.
- When you upgrade from STAR Reading version 3 to version 4 or higher, all shortcuts to the student program will automatically redirect to the browser-based program (the Renaissance Place Welcome page) each time they are used.

Improvements Specific to STAR Reading Version 4.3 RP

STAR Reading versions 3 RP to 4.2 RP were identical in content to STAR Reading version 2. Changes in content were made for version 4.3 RP, along with other changes, all described below.

- The item bank was further expanded. A total of 626 new items were added, and several hundred were retired. The resulting STAR Reading 4.3 RP item bank had 1,792 items graded into 54 difficulty levels.
- The Adaptive Branching process was further improved by changing the difficulty target used to select each item. The new difficulty target further improves the measurement efficiency of STAR Reading, and is expected to increase measurement precision, score reliability, and test validity.
- A new feature, dynamic calibration, was added. Dynamic calibration makes it possible to include small numbers of unscored items in selected students' tests, for the purpose of collecting item response data for research and development use.
- STAR Reading can now be used to test kindergarten students, at the teacher's discretion. Kindergartners' score reports will include Scaled Scores (SS), Instructional Reading Levels (IRL), Grade Equivalent (GE) scores, and Zone of Proximal Development (ZPD) ranges, but not such norm-referenced scores as Percentile Ranks (PR) and Normal Curve Equivalent (NCE) scores.

Improvements Specific to STAR Reading Version 4.4 RP

- For STAR Reading 4.4, 285 new test items were added and 29 items were retired, for a total of 2,048 test items.
- Test items could be re-used after 90 days, allowing for more frequent testing, if desired. (Prior to version 4.4, items would not be re-used for 180 days.)

Improvements Specific to STAR Reading Enterprise ENTERPRISE

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

- The item bank was expanded to support STAR Reading Enterprise. In addition to 2,125 items in the original vocabulary-in-context format, and 672 longer, authentic passage comprehension items, the item bank now includes more than 2,100 items measuring the new domains, skill sets, and specific skills that distinguish Enterprise from STAR Reading Service.

Test Interface

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

- If using the keyboard, students press one of the four letter keys (**A**, **B**, **C**, and **D**) and then press the **Enter** key (or the **return** key on Macintosh computers).
- If using the mouse, students click the answer of choice and then click **Next** to enter the answer.

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

Practice Session

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

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

Adaptive Branching/Test Length

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

In order to minimize student frustration, the first administration of the STAR Reading test begins with items that have a difficulty level that is below what a typical student at a given grade can handle—usually one or two grades below grade placement. On the average, about 86 percent of students will be able to answer the first item correctly. Teachers can override this typical value by entering an even lower Estimated Instructional Reading Level for the student. On the second and subsequent administrations, the STAR Reading test again begins with items that have a difficulty level lower than the previously demonstrated reading ability. Students generally have an 85 percent chance of answering the first item correctly on second and subsequent tests.

Test Length: STAR Reading

Once the testing session is underway, the STAR Reading test administers 25 items (the STAR Reading Enterprise test administers 34 items) of varying difficulty based on the student's responses; this is sufficient information to obtain a reliable Scaled Score and to determine the student's Instructional Reading Level. The length of time needed to complete a STAR Reading test varies across students. Table 1 provides an overview of the testing time by grade for the students participating in the spring 2008 norming study (see the chapter on norming, page 112, for more information). The results of the analysis of test completion time indicates that about half of the students at every grade will complete the STAR Reading test in less than 9 minutes, and even in the slowest grade (grade 1) 95 percent of students finished their STAR Reading test in less than 15 minutes.

Table 1: Percentiles of Total Time to Complete STAR Reading Assessment in the 2008 Norming Study

	Time to Complete Test (in Minutes)				
Grade	**5th Percentile**	**25th Percentile**	**50th Percentile**	**75th Percentile**	**95th Percentile**
1	5.2	7.2	8.9	10.9	14.3
2	4.7	6.3	7.8	9.6	12.7
3	5.1	7.0	8.4	10.1	12.8
4	4.8	6.5	8.0	9.6	12.2
5	4.6	6.3	7.7	9.3	12.0
6	4.5	6.1	7.5	9.1	11.8
7	4.3	5.8	7.2	8.8	11.4
8	4.0	5.8	7.1	8.7	11.3
9	3.9	5.5	6.8	8.3	11.3
10	4.0	5.8	7.1	8.7	11.2
11	3.9	5.6	6.9	8.6	11.1
12	3.9	5.6	7.0	8.6	11.0

Test Length: STAR Reading Enterprise ENTERPRISE

Once the testing session is underway, STAR Reading Enterprise administers up to 5 practice items, plus 34 items of varying difficulty based on the student's responses. This is sufficient information to obtain a reliable Scaled Score. The length of time needed to complete a STAR Reading Enterprise test varies across students. Table 2 provides an overview of the testing time by grade for the students who took STAR Reading Enterprise during the first three months following its release in the summer of 2011. The results of the analysis of test completion time indicates that half or more of students will complete the test in 11–18 minutes, depending on grade, and even in the slowest grade (grade 3) 95% of students finished their STAR Reading Enterprise test in less than 28 minutes.

Table 2: Percentiles of Total Time to Complete STAR Reading Enterprise Test Items During Its First Three Months of Operational Use, July–September 2011

Grade	Number of Tests	Time in Minutes				
		5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile
K	2,678	4.4	7.6	11.0	15.4	22.1
1	100,149	4.0	7.3	11.2	16.3	24.0
2	231,745	5.8	11.4	15.6	19.3	24.4
3	252,851	7.6	13.6	17.6	21.6	27.6
4	243,363	8.4	13.9	17.5	21.1	26.7
5	238,681	8.8	13.7	16.9	20.3	25.4
6	177,454	8.9	13.6	16.7	19.9	24.9
7	132,765	8.2	12.4	15.3	18.5	23.4
8	126,952	8.0	12.1	14.9	17.9	22.7
9	59,104	8.0	12.3	15.2	18.4	23.4
10	42,541	7.7	12.0	15.0	18.2	23.2
11	27,671	7.6	12.1	15.0	18.2	23.3
12	21,525	7.4	11.9	14.9	18.2	23.3

Test Repetition

STAR Reading 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.

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

Item Time Limits

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

Table 3: STAR Reading Time-Out Limits

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

a. Vocabulary-in-context items.

b. Authentic text/passage comprehension items.

These time-out values are based on latency data obtained during item validation. Very few vocabulary-in-context items at any grade had latencies longer than 30 seconds, and almost none (fewer than 0.3 percent) had latencies of more than 45 seconds. Thus, the time-out limit was set to 45 seconds for most students and increased to 60 seconds for the very young students.

ENTERPRISE Table 4 shows time limits for STAR Reading Enterprise test questions:

Table 4: STAR Reading Enterprise Time-Out Limits

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

a. Vocabulary-in-context items.

b. Items from 5 domains in 5 blocks, including some vocabulary-in-context.

c. 60 seconds for vocabulary-in-context items.

d. 180 seconds for vocabulary-in-context items.

e. 45 seconds for vocabulary-in-context items.

f. 135 seconds for vocabulary-in-context items.

Beginning with version 2.2, STAR Reading provides the option of extended time limits for selected students who, in the judgment of the test administrator, require more than the standard amount of time to read and answer the test questions. Extended time may be a valuable accommodation for English language learners as well as for some students with disabilities. Test users who elect the extended time limit for their students should be aware that STAR Reading norms, as well as other technical data such as reliability and validity, are based on test administration using the standard time limits. When the extended time limit accommodation is elected, students have three times longer than the standard time limits to answer each question.

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

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

Test Security

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

Split-Application Model

In the STAR Reading RP software, when students log in, they do not have access to the same functions that teachers, administrators, and other personnel can access. Students are allowed to test, but they have no other tasks available in STAR Reading RP; 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 Reading test consists of items chosen from a large number of items of similar difficulty based on the student's estimated ability. Because each test is individually assembled based on the student's past and present performance, identical sequences of items are rare. This feature, while motivated chiefly by psychometric considerations, contributes to test security by limiting the impact of item exposure.

Data Encryption

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

Access Levels and Capabilities

Each user’s level of access to a Renaissance Place program depends on the primary position assigned to that user and the capabilities the user has been granted in the Renaissance Place program. Each primary position is part of a user group. There are seven user groups: district administrator, district staff, school administrator, school staff, teacher, parent, and student. By default, each user group is granted a specific set of capabilities. Each capability corresponds to one or more tasks that can be performed in the program. The capabilities in these sets can be changed; capabilities can also be granted or removed on an individual level. Since users can be assigned to the district and/or one or more schools (and be assigned different primary positions at the different locations), and since the capabilities granted to a user can be customized, there are many levels of access an individual user can have.

Renaissance Place also allows you to restrict students’ access to certain computers. This prevents students from taking STAR Reading RP tests from unauthorized computers (such as home computers). For more information on access and security, see the *Renaissance Place Software Manual.*

The security of the STAR Reading RP 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, or staff) must enter a user name and password before they can access the data and create reports. Parents who are granted access to Renaissance Place must also log in with a user name and password before they can access the Parent Report. Without an appropriate user name and password, personnel and parents cannot use the STAR Reading RP software.

Test Monitoring/Password Entry

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

Final Caveat

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

STAR Reading Enterprise and the Common Core State Standards ENTERPRISE

The Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects are based on an integrated model of literacy, but are divided into four areas of literacy to provide clarity for student expectations in reading, writing, speaking, and listening. Reading is further divided into three areas: Foundational Skills K–5, Reading Standards for Literature K–12, and the Reading Standards for Informational Text K–12. Each area of literacy has a set of consistent anchor standards that define the general literacy expectations for students to be college and career ready. The ten anchor standards for literature and informational text are the same and are organized into four groupings:

- Key Ideas and Details
- Craft and Structure
- Integration of Knowledge and Ideas
- Range of Reading and Level of Text Complexity

Each grade level K–8 and the high school grade bands of 9–10 and 11–12 have grade-specific standards based in the anchor expectations for that grade level. The reading standards stress the importance of both the complexity of text that students read and the skills they use to read. The grade-by-grade expectations delineate steady growth in text complexity and skills acquisition resulting in increasingly sophisticated understanding and use of the text. The Reading Standards for Literature K–5 and the Reading Standards for Informational Text K–5 are the pertinent areas of literacy for STAR Reading Enterprise.

STAR Reading Enterprise is a K–12 assessment that focuses on measuring student performance with skills in five domains:

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

Specific grade-level expectations are identified in each domain. Measures in these areas provide valuable information regarding the acquisition of reading ability along the continuum of literacy expectations. Resources consulted to determine the set of skills most appropriate for assessing reading development include:

- Reading Next—A Vision for Action and Research in Middle and High School Literacy: A Report to Carnegie Corporation of New York. © 2004 by Carnegie Corporation of New York. http://www.all4ed.org/files/ReadingNext.pdf.
- NCTE Principles of Adolescent Literacy Reform, A Policy Research Brief, Produced by The National Council of Teachers of English, April 2006. http://www.ncte.org/library/NCTEFiles/Resources/Positions/Adol-Lit-Brief.pdf.
- Improving Adolescent Literacy: Effective Classroom and Intervention Practices, August 2008. http://eric.ed.gov/PDFS/ED502398.pdf.
- Reading Framework for the 2009 National Assessment of Education Progress. http://www.nagb.org/publications/frameworks/reading09.pdf.
- Common Core State Standards Initiative (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects.
- Thomas B. Fordham Institute's study, The State of State Standards—and the Common Core—in 2010.
- Experts in the field of reading instruction and assessment.
- Exemplary state standards.

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

The Common Core State Standards Initiative recognizes that students should "read widely and deeply from among a broad range of high-quality, increasingly challenging literary and informational texts" and that "students must also show a steadily growing ability to discern more from and make fuller use of text, including making an increasing number of connections among ideas and between texts, considering a wider range of textual evidence, and becoming more sensitive to inconsistencies, ambiguities, and poor reasoning in texts" (Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects 2010). The standards provide grade-level specific standards that delineate the progress toward these goals.

Core Progress for reading, a research-based and empirically supported learning progression of reading, identifies the continuum of reading strategies, behaviors, and skills 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 Reading Enterprise are a subset of this larger continuum of skills. STAR Reading Enterprise assessment results are correlated to the Core Progress learning progression for reading.

Test Administration Procedures

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

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

Content and Item Development

Content Specification: STAR Reading

The content of STAR Reading 2 is identical to the content in versions 3 RP. Content development was driven by the test design and test purposes, which are to measure comprehension and general reading achievement. Based on test purpose, the desired content had to meet certain criteria. First, it had to cover a range broad enough to test students from grades K–12. Thus, items had to represent reading levels ranging all the way from kindergarten through post-high school. Second, the final collection of test items had to be large enough so that students could test up to five times per year without being given the same items twice.

The current item bank for STAR Reading 4.4 contains a total of 2,048 items: 1,620 vocabulary-in-context items, and 428 authentic text passage items.

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

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

Beginning with new test items introduced in version 4.3, STAR Reading item developers used ATOS instead of the EDL word list. ATOS is a system for evaluating the reading level of continuous text; it contains 23,000 words in its graded vocabulary list. This readability formula was developed by Renaissance Learning, Inc., and designed by leading readability experts. ATOS is the first formula to include statistics from actual student book reading (over 30,000 students, reading almost 1,000,000 books).

Content Specification: STAR Reading Enterprise ENTERPRISE

STAR Reading Enterprise is an expanded test with new content and several technical innovations. STAR Reading Enterprise consists of 4,156 operational items that align to a set of reading skills derived from exemplary state standards as well as CCSS and current research. The items are intended to measure progress in reading skills as defined by the learning progression for reading. Core progress learning progression for reading consists of 36 skills organized within 5 domains of reading (see Table 5), and maps the progressions of reading skills and understandings as they develop in sophistication from kindergarten through grade 12. Each STAR item is designed to assess a specific skill within the progression. For more information on Core Progress for reading, refer to the white paper, Core Progress for Reading: An empirically validated learning progression (http://doc.renlearn.com/KMNet/R0053985FA6D567F.pdf).

For information regarding the development of STAR Reading items, see "Item Development Specifications: STAR Reading" on page 23. Before inclusion in the STAR Reading Enterprise item bank, all STAR Reading items were reviewed to ensure they met the content specifications for STAR Reading Enterprise item development. Items that did not meet the specifications were revised and recalibrated. All new item development adheres to the content specifications.

All items were calibrated using the dynamic calibration method. The first stage of the expanded STAR Reading Enterprise development was identifying the set of skills to be assessed. Multiple resources were consulted to determine the set of skills most appropriate for assessing the reading development of K–12 US students. The resources include but are not limited to:

- Reading Next—A Vision for Action and Research in Middle and High School Literacy: A Report to Carnegie Corporation of New York © 2004 by Carnegie Corporation of New York. http://www.all4ed.org/files/ReadingNext.pdf.
- NCTE Principles of Adolescent Literacy Reform, A Policy Research Brief, Produced by The National Council of Teachers of English, April 2006. http://www.ncte.org/library/NCTEFiles/Resources/Positions/Adol-Lit-Brief.pdf.
- Improving Adolescent Literacy: Effective Classroom and Intervention Practices, August 2008. http://eric.ed.gov/PDFS/ED502398.pdf.
- Reading Framework for the 2009 National Assessment of Education Progress. http://www.nagb.org/publications/frameworks/reading09.pdf.
- Common Core State Standards Initiative (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects.

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

- Word Knowledge and Skills
- Comprehension Strategies and Constructing Meaning
- Analyzing Literary Text

- Understanding Author’s Craft
- Analyzing Argument and Evaluating Text

Table 5: Core Progress for Reading: Domains and Skills

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

An Example of STAR Reading Enterprise Item Adherence to a Specific Skill within Core Progress for Reading

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

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

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

Item Development Specifications: STAR Reading

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

Vocabulary-in-Context Item Specifications

Once the test design was determined, individual test items were assembled for tryout and calibration. For the STAR Reading 2 test, the item tryout and calibration included all 838 vocabulary items from the STAR Reading 1 test, plus 836 new vocabulary items created for the STAR Reading 2 test. Because all items go through a rigorous calibration process, approximately 100 new questions at each grade level were written and calibrated to ensure that approximately 60 new items per level would be acceptable for the final item collection. (Due to the limited number of primer words available for the kindergarten level, the starting set for this level contained only 30 items.) Having a pool of almost 1,700 vocabulary items allowed significant flexibility in selecting only the best items from each group for the final product.

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

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

2. To answer the question, the student selects the word from the answer choices that best completes the sentence. The correct answer option is the word that appropriately fits both the semantics and the syntax of the sentence. All of the incorrect answer options either fit the syntax of the sentence or relate to the meaning of something in the sentence. They do not, however, meet both conditions.

3. The answer blanks are generally located near the end of the context sentence to minimize the amount of rereading required.

4. The sentence provides sufficient context clues for students to determine the appropriate answer choice. However, the length of each sentence varies according to the guidelines shown in Table 6.

5. Typically, the words that provide the context clues in the sentence are below the level of the actual test word. However, due to a limited number of available words, not all of the questions at or below grade 2 meet this criterion—but even at these levels, no context words are above the grade level of the item.

6. The correct answer option is a word selected from the appropriate grade level of the item set. Incorrect answer choices are words at the same test level or one grade below. Through vocabulary-in-context test items, STAR Reading requires students to rely on background information, apply vocabulary knowledge, and use active strategies to construct meaning from the assessment text. These cognitive tasks are consistent with what researchers and practitioners describe as reading comprehension.

Table 6: Maximum Sentence Length per Item Grade Level

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

Authentic Text Passage Item Specifications[1]

STAR Reading 2 and higher authentic text passage items are passages of extended text administered to students at grade levels 3–13. These items were developed by identifying authentic texts, extracting appropriate passages, and creating cloze-type questions and answers. Each passage is comprised of content that can stand alone as a unified, coherent text. Items were selected which assess passage-level, not merely sentence-level, understanding. To answer the item correctly, the student needs to have a general understanding of the context and content of the passage, not merely an understanding of the specific content of the sentence.

The first authentic passages in STAR Reading were extracted from children's and young adult literature, from nonfiction books, and from newspapers, magazines, and encyclopedias. Passages were selected from combinations of three primary

[1] At the earliest grade levels (K–3), some original passages were written. This was due to the dearth of Accelerated Reader texts available at the beginning reading levels.

categories for school-age children: popular fiction, classic fiction, and nonfiction. Overall Flesch-Kincaid readability estimates of the source materials were used as initial estimates of grade-level difficulty.

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

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

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

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

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

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

Consideration was given to include some passages from the public domain. When necessary, original long items were written. In any case, passages excerpted or adapted are attributed in "Item and Scale Calibration" on page 32.

The STAR Reading 2 item tryout and calibration included 459 authentic text passage items. About 40 questions at each grade level from 3–13 were tested to ensure that approximately 25 items per level would be acceptable for the final item collection. (No authentic text passage items were developed for grade levels 1 and 2, as the STAR Reading 2 design called solely for the use of shorter vocabulary-in-context items at those two grade levels.)

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

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

Table 7: Authentic Text Passage Length

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

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

Item Development Specifications: STAR Reading Enterprise ENTERPRISE

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

Adherence to Skills

STAR Reading Enterprise assesses more than 470 grade-specific skills within Core Progress learning progression for reading. Item development is skill-specific. Each item in the item bank is developed for and clearly aligned to one skill. An item meets the alignment criteria if the knowledge and skill required to correctly answer the item match the intended knowledge and skill. Answering an item correctly does not require reading skill knowledge beyond the expected knowledge for the skill being assessed. STAR Reading items include only the information and text needed to assess the skill.

Level of Difficulty: Readability

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

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

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

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

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

Table 8: Readability Guidelines Table

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

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

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

- Cognitive Load: Cognitive load involves the type and amount of knowledge and thinking that a student must have and use in order to answer the item correctly. The entire impact of the stem and answer choices must be taken into account.

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

Efficiency in Use of Student Time

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

Balanced Items: Bias and Fairness

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

- Items are free of stereotyping, representing different groups of people in non-stereotypical settings.
- Items do not refer to inappropriate content that includes, but is not limited to content that presents stereotypes based on ethnicity, gender, culture, economic class, or religion.
- Items do not present any ethnicity, gender, culture, economic class, or religion unfavorably.
- Items do not introduce inappropriate information, settings, or situations.
- Items do not reference illegal activities, sinister or depressing subjects, religious activities or holidays based on religious activities, witchcraft, or unsafe activities.

Accuracy of Content

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

Language Conventions

Grammar, usage, mechanics, and spelling conventions in all STAR Reading items adhere to the rules and guidelines in the approved content reference books. *Merriam Webster's 11th Edition* is the reference for pronunciation and spelling. *The Chicago Manual of Style 16th Edition* and *The Little, Brown Handbook* are the anchor references for grammar, mechanics, and usage.

Item Components

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

Item stems meet the following criteria with limited exceptions:

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

Distractors meet the following criteria with limited exceptions:

- All distractors are plausible and reasonable.
- Distractors do not contain clues that hint at correct or incorrect distractors. Incorrect answers are created based on common student mistakes.

- Distractors that are not common mistakes may vary between being close to the correct answer or close to a distractor that is the result of a common mistake.
- Distractors are independent of each other, are approximately the same length, have grammatically parallel structure, and are grammatically consistent with the stem.
- *None of these, none of the above, not given, all of the above,* and *all of these* are not used as distractors.

Item and Scale Calibration

Background

The introduction of STAR Reading Enterprise marks a major evolution in the calibration of STAR Reading items. For the original versions of STAR Reading, from version 1 (circa 1995) and all versions prior to version 4.3, data for item calibration were collected using printed test booklets and answer sheets, in which the items were formatted to closely match the appearance those items would later take when displayed on computer screens. For STAR Reading versions 4.3 and later, and for STAR Reading Enterprise, new test items to be calibrated were embedded as unscored items in STAR Reading itself, and the data for calibration were collected by the STAR Reading software. Renaissance Learning calls this data collection process *dynamic calibration.*

The dynamic calibration feature 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 toward 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 do not count toward students' final scores, but will be analyzed in conjunction with the responses of thousands of other students.

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

The item bank used in version 1 of STAR Reading contained 838 vocabulary in context items. STAR Reading version 2 had 1,409 test items, including both vocabulary in context and authentic text passage items. In STAR Reading version 4.3 RP, the adaptive test item bank consisted of 1,792 calibrated test items. Of these, 626 items were new, and 1,166 items were carried over from the set of 1,409 test items that were developed for use in STAR Reading version 2 and used in that and later versions up to and including version 4.1 RP. In STAR Reading version 4.4 RP, 285 new items were added to the test bank and 29 items were retired, resulting in an item bank totaling 2,048 test items.

Items carried over from version 2 had been calibrated by administering them to national student samples in printed test booklets. Items developed specifically for version 4.3 and above were calibrated online, by using the dynamic calibration feature to embed them in otherwise normal STAR Reading tests. This chapter describes both booklet-based and dynamic item calibration efforts, in order to provide a comprehensive summary of the technical approaches used to calibrate STAR Reading test items from version 2 through the present.

Calibration of STAR Reading Items for Use in Version 2

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

Regardless of how carefully test items are written and edited, it is critical to study how students actually perform on each item. The first large-scale research activity undertaken in creating the test was the item validation program conducted in March 1995. This project provided data concerning the technical and statistical quality of each test item written for the STAR Reading test. The results of the item validation study were used to decide whether item grade assignments, or “tags,” were correct as obtained from the EDL vocabulary list, or whether they needed to be adjusted up or down based on student response data. This refinement of the item grade level tags made the STAR Reading criterion reference more timely.

In STAR Reading 2 development, a large-scale item calibration program was conducted in the spring of 1998. The STAR Reading 2 item calibration study incorporated all of the newly written vocabulary-in-context and authentic text passage items, as well as all 838 vocabulary items in the STAR Reading 1 item bank. Two distinct phases comprised the item calibration study. The first phase was the collection of item response data from a multi-level national student sample. The second phase involved the fitting of item response models to the data, and developing a single IRT difficulty scale spanning all levels from grades 1–12.

Sample Description

The data collection phase of the STAR Reading 2 calibration study began with a total item pool of 2,133 items. A nationally representative sample of students tested these items. A total of 27,807 students from 247 schools participated in the item calibration study. Table 9 provides the numbers of students in each grade who participated in the study.

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

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

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

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

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

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

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

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

Item Presentation

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

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

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

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

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

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

The number of field test forms constructed for each of the seven test levels is shown in the last column of Table 12 (varying from 11–15 forms per level). Calibration test forms were spiraled within a classroom such that each student received a test form essentially at random. This design ensured that no more than two or three students in any classroom attempted any particular tryout item. Additionally, it ensured a balance of student ability across the various tryout forms. Typically, 250–300 students at the designated grade level of the test item received a given question on their test.

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

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

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

Item Difficulty

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

Item Discrimination

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

Item Response Function

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

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

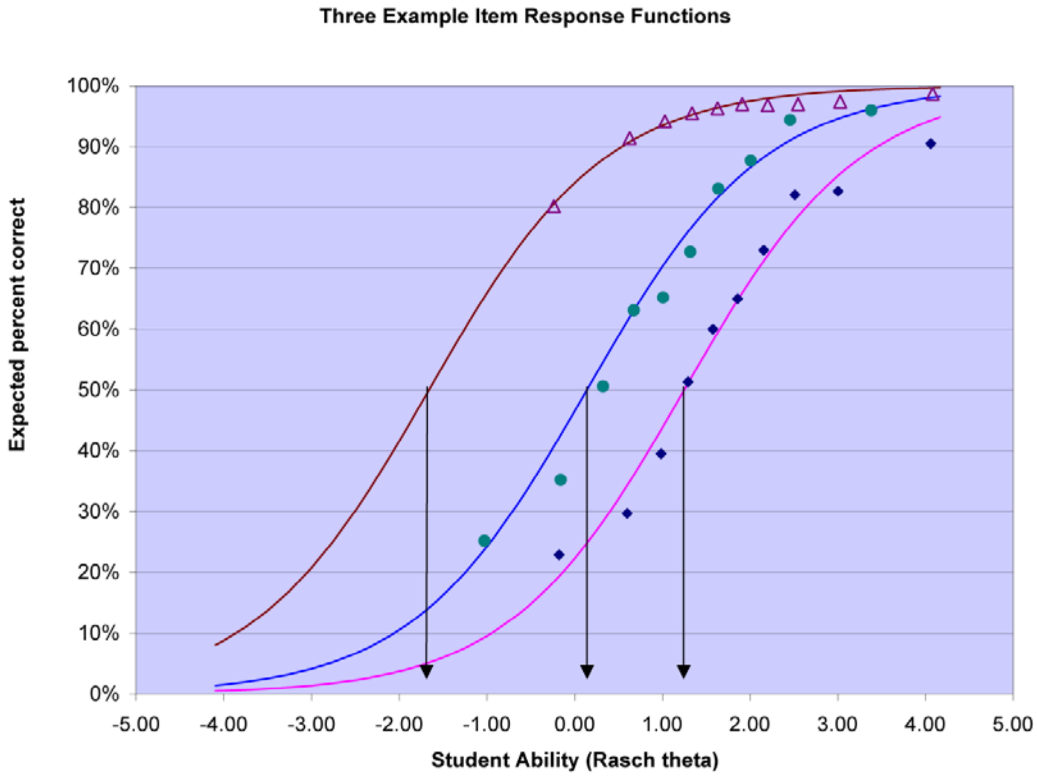
Figure 1 is a plot of three item response functions: one for an easy item, one for a more difficult one, and one for a very difficult item. Each plot is a continuous S-shaped (ogive) curve. The horizontal axis is the scale of student ability, ranging from very low ability (–5.0 on the scale) to very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

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

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

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

Figure 1: Example of Item Statistics Database Presentation of Information



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

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

Rules for Item Retention

Following these analyses, each test item, along with both traditional and IRT analysis information (including IRF and EIRF plots) and information about the test level, form, and item identifier, were stored in an item statistics database. A panel of content reviewers then examined each item, within content strands, to determine whether the item met all criteria for inclusion into the bank of items that would be used in the norming version of the STAR Reading 2 test. The item statistics database allowed experts easy access to all available information about an item in order to interactively designate items that, in their opinion, did not meet acceptable standards for inclusion in the STAR Reading 2 item bank.

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

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

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

Of the initial 2,133 items administered in the STAR Reading 2 calibration research study, 1,409 were deemed of sufficient quality to be retained for further analyses. Traditional item-level analyses were conducted again on the reduced data set that excluded the eliminated items. IRT calibration was also performed on the reduced data set and all test forms and levels were equated based on the information provided by the embedded anchor items within each test form. This resulted in placing the IRT item difficulty parameters for all items onto a single scale spanning grades 1–12.

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

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

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

Calibration of Supplemental Items for Use in Version 4.3 RP

For version 4.3 RP, 854 new test items were written for the purpose of replacing some of the items originally in version 2, and supplementing the item bank as well. These items were written to the same content specifications used to develop STAR Reading version 2 items.

To calibrate these new items, processes quite different from those used with version 2.0 were used. Data collection was accomplished by using dynamic calibration to embed between three and five new items at random points in STAR Reading 4.1 RP tests; responses to the new items were not used in scoring students' performance on the STAR Reading tests. Data collection took place in "hosted" Renaissance Place (RP) sites—that is, schools whose STAR Reading tests were administered by accessing remote network servers operated by Renaissance Learning. Over a 10-week period between September and November 2007, nearly a million students in grades 1–12 in 48 states participated in this data collection effort. The effort took place in two overlapping batches of new test items, with the first batch used in the first half of the data collection period, and the second batch used during the second half.

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

An average of over 4,000 students responded to each new test item. These item responses were analyzed using two different methods to calibrate each item. The first method employed the Winsteps Rasch calibration software. For the Winsteps analyses, student Rasch ability estimates were fixed at the values calculated during the STAR Reading tests, and the Rasch difficulty parameters were estimated from analyses of the response data and the fixed ability estimates. The second method used the SAS/STAT™ software to estimate the threshold (difficulty) parameter of every new item by calculating the non-linear regression of each new item score (0 or 1) on the STAR Reading Rasch ability estimates. The purpose of employing these two different difficulty calibration methods was to corroborate the accuracy of the non-linear regression approach. The two methods yielded virtually identical results.

The Winsteps analysis also produced the same fit indices used during the calibration of STAR Reading 2 items. Those fit indices, along with the proportion correct and item test score correlation statistics, were used as the basis for item retention decisions, with criteria similar to those applied during development of the STAR Reading 2 item bank, summarized earlier. Applying these criteria resulted in the retention of 626 of the 854 new test items for use in the version 4.3 RP item bank. These 626 newly calibrated items were used to replace 243 of the old items, and to supplement the rest of the item bank. The resulting version 4.3 RP item bank consisted of 1,792 test items.

Table 14 summarizes the final analysis information for the new test items, overall and by the target grades tagged to each item. The data in Table 13 can be compared with those in Table 14 to compare the STAR Reading 2 and 4.3 RP item analysis results.

Table 14: Calibration Test Item Summary Information by Test Item Grade Level, STAR Reading 4.3 Calibration Study–Fall 2007

Item Grade Level	Number of Items	Sample Size[a]	Average Percent Correct	Median Percent Correct	Median Point-Biserial	Average Scaled Difficulty	Average Scaled Ability
K	51	230,580	78	78	47	–3.77	–1.65
1	68	238,578	82	82	45	–3.68	–1.23
2	99	460,175	76	76	51	–2.91	–1.06
3	130	693,184	74	78	47	–1.91	–0.23
4	69	543,554	74	78	41	–1.05	0.64
5	44	514,146	70	72	40	–0.14	1.24
6	32	321,855	71	72	38	0.15	1.62
7	42	402,530	60	58	37	1.40	2.07

Table 14: Calibration Test Item Summary Information by Test Item Grade Level, STAR Reading 4.3 Calibration Study–Fall 2007 (Continued)

Item Grade Level	Number of Items	Sample Size[a]	Average Percent Correct	Median Percent Correct	Median Point-Biserial	Average Scaled Difficulty	Average Scaled Ability
8	46	317,110	55	53	33	2.10	2.36
9	36	174,906	54	50	33	2.39	2.59
10	56	99,387	51	54	31	2.95	2.91
11	68	62,596	47	43	22	3.50	3.12
12	51	43,343	44	41	18	3.60	3.11
> 12	62	52,359	34	31	11	4.30	3.10

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

Computer-Adaptive Test Design

The third phase of content specification is determined by the student's performance during testing. In the conventional paper-and-pencil standardized test, items retained from the item tryout or item calibration study are organized by level; then, each student takes all items within a given test level. Thus, the student is only tested on reading skills deemed to be appropriate for his or her grade level. In computer-adaptive tests like the STAR Reading test, the items taken by a student are dynamically selected in light of that student's performance during the testing session. Thus, a low-performing student's reading skills may branch to easier items in order to better estimate his or her reading achievement level. High-performing students may branch to more challenging reading items in order to better determine the breadth of their reading skills and their reading achievement level.

Items retained from the STAR Reading item calibration studies have been organized into two large item "pools" (vocabulary-in-context items and authentic text passage items), each ordered from the easiest to most difficult. During an adaptive test, a student may be "routed" to items at the lowest reading level or to items at higher reading levels within the overall pool of items, depending on the student's unfolding performance during the testing session. In general, when an item is answered correctly, the student is then given a more difficult item. When an item is answered incorrectly, the student is then given an easier item. Item difficulty here is defined by results of the STAR Reading item calibration studies.

All STAR Reading tests between version 2 and 4.3 RP, inclusive, administer a fixed-length, 25-item, computer-adaptive test. Students who have not taken a

STAR Reading test within six months initially receive an item whose difficulty level is relatively easy for students at that grade level. The selection of an item that is a bit easier than average minimizes any effects of initial anxiety that students may have when starting the test and serves to better facilitate the student's initial reactions to the test. These starting points vary by grade level and were based on research conducted as part of the national item calibration study.

When a student has taken a STAR Reading test within the last six months, the difficulty of the first item depends on that student's previous STAR Reading test score information. After the administration of the initial item, and after the student has entered an answer, STAR Reading software estimates the student's reading ability. The software then selects the next item randomly from among all of the items available that closely match the student's estimated reading ability.

Randomization of items with difficulty values near the student's adjusted reading ability allows the program to avoid overexposure of test items. All items in grades K–2 tests, and the first twenty items in grade 3–12 tests, are dynamically selected from an item bank consisting of all the retained vocabulary-in-context items. For grades 3–12, the second part of the test (the last five items) begins once a good estimate of the student's reading ability has been established and then selects items from a pool of authentic text passage items to refine the student's final estimated reading ability. Items that have been administered to the same student within the past three-month time period are not available for administration. The large numbers of items available in the item pools, however, ensure that this minor constraint has negligible impact on the quality of each STAR Reading RP computer-adaptive test.

Scoring in the STAR Reading Tests

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

This approach to scoring enables the STAR Reading 3 RP and higher test to provide Scaled Scores that are statistically consistent and efficient. Accompanying each Scaled Score is an associated measure of the degree of uncertainty, called the conditional standard error of measurement (CSEM). Unlike a conventional paper-and-pencil test, the CSEM values for the STAR Reading test are unique for each student. CSEM values are dependent on the particular items the student received and on the student's performance on those items.

Scaled Scores are expressed on a common scale that spans all grade levels covered by the STAR Reading 3 RP and higher test (grades K–12). Because of this common scale, Scaled Scores are directly comparable with each other, regardless of grade level. Other scores, such as Percentile Ranks and Grade Equivalents, are derived from the Scaled Scores obtained in the STAR Reading norming study described in the "Norming" section of this manual.

Scale Calibration

The outcome of the first item calibration study described above was a sizable bank of test items suitable for use in the STAR Reading 2 test, with an IRT difficulty scale parameter for each item. The second calibration study yielded an additional 626 calibrated items. The item difficulty scale itself was devised such that it spanned a range of item difficulty from grades 1–12. An important feature of Item Response Theory is that the same scale used to characterize the difficulty of the test items is also used to characterize examinees' ability; in fact, IRT models express the probability of a correct response as a function of the difference between the scale values of an item's difficulty and an examinee's ability. The IRT ability/difficulty scale is continuous; in the STAR Reading norming studies described in the "Norming" section, the values of observed ability ranged from about –7.3 to +9.2, with the zero value occurring at about the sixth-grade level.

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

The Linking Study

From around the country and spanning all 12 grades, 4,589 students participated in the linking study. Linking study participants took both STAR Reading 1 and

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

Summary statistics of the test scores of the 4,551 cases included in the linking analysis are listed in Table 15. The table lists actual STAR Reading 1 Scaled Score means and standard deviations, as well as the same statistics for STAR Reading 2 IRT ability estimates and equivalent Scaled Scores calculated using the conversion table from the linking study. Comparing the STAR Reading 1 Scaled Score means to the IRT ability score means illustrates how different the two metrics are. Comparing the STAR Reading 1 Scaled Score means to the STAR Reading 2 Equivalent Scale Scores in the rightmost two columns of Table 15 illustrates how successful the scale linking was.

Table 15: Summary Statistics of STAR Reading 1 and 2 Scores from the Linking Study, by Grade—Spring 1999 (N = 4,551 Students)

Grade Level	Sample Size	STAR Reading 1 Scaled Scores		STAR Reading 2 IRT Ability Scores		STAR Reading 2 Equivalent Scale Scores	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
1	284	216	95	–1.98	1.48	208	109
2	772	339	115	–0.43	1.60	344	148
3	476	419	128	0.33	1.53	419	153
4	554	490	152	0.91	1.51	490	187
5	520	652	176	2.12	1.31	661	213
6	219	785	222	2.98	1.29	823	248
7	702	946	228	3.57	1.18	943	247
8	545	958	285	3.64	1.40	963	276
9	179	967	301	3.51	1.59	942	292
10	81	1,079	292	4.03	1.81	1,047	323

Table 15: Summary Statistics of STAR Reading 1 and 2 Scores from the Linking Study, by Grade—Spring 1999 (N = 4,551 Students) (Continued)

Grade Level	Sample Size	STAR Reading 1 Scaled Scores		STAR Reading 2 IRT Ability Scores		STAR Reading 2 Equivalent Scale Scores	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
11	156	1,031	310	3.98	1.53	1,024	287
12	63	1,157	299	4.81	1.42	1,169	229
1–12	4,551	656	345	1.73	2.36	658	353

Data from the linking study made it clear that STAR Reading 2 software measures ability levels extending beyond the minimum and maximum STAR Reading 1 Scaled Scores. In order to retain the superior bandwidth of STAR Reading 2 software, extrapolation procedures were used to extend the Scaled Score range below 50 and above 1,350.

Table 16 contains an excerpt from the IRT ability to Scaled Score conversion table that was developed in the course of the linking study.

Table 16: Example IRT Ability to Equivalent Scaled Score Conversion

IRT Ability		Equivalent Scaled Score
From	To	
–6.2845	–6.2430	50
–3.1790	–3.1525	100
–2.5030	–2.4910	150
–1.9030	–1.8910	200
–1.2955	–1.2840	250
–0.7075	–0.6980	300
–0.1805	–0.1715	350
0.3390	0.3490	400
0.7600	0.7695	450
1.2450	1.2550	500
1.6205	1.6270	550
1.9990	2.0045	600
2.3240	2.3300	650
2.5985	2.6030	700
2.8160	2.8185	750

Table 16: Example IRT Ability to Equivalent Scaled Score Conversion (Continued)

IRT Ability		
From	**To**	**Equivalent Scaled Score**
3.0090	3.0130	800
3.2120	3.2180	850
3.4570	3.4635	900
3.7435	3.7485	950
3.9560	3.9580	1,000
4.2120	4.2165	1,100
4.3645	4.3680	1,150
4.5785	4.5820	1,200
4.8280	4.8345	1,250
5.0940	5.1020	1,300
7.5920	7.6340	1,350
9.6870 and above		1,400

Calibration of STAR Reading Skills Items for Use in STAR Reading Enterprise ENTERPRISE

For the development of STAR Reading Enterprise, several thousand new items spanning content appropriate for grades 1–12 were developed. Unlike previous versions of STAR Reading, which were designed to measure only reading comprehension, STAR Reading Enterprise items were designed to measure dozens of discrete reading skills. Data for calibrating them were collected using the dynamic calibration feature of the Renaissance Place versions of STAR Reading. Small numbers of these items were randomly selected for each student, and administered in a separate reading skills section administered at the end of the regular STAR Reading test. The reading skills section was administered using Renaissance Place, beginning in the 2008–2009 school year, and continuing to the present. Each student taking STAR Reading on Renaissance Place was administered a small number of these new, uncalibrated items.

Reliability and Measurement Precision

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

Second, reliability as a measure of consistency between two different measurements indicates the extent to which a test yields consistent results from one administration to another and from one test form to another. Tests must yield somewhat consistent results in order to be useful; the reliability co-efficient is obtained by calculating the correlation between students' scores on two different occasions, or on two alternate versions of the test given at the same occasion. Because the amount of the attribute being measured may change over time, and the content of tests may differ from one version to another, a test with 95 percent internal consistency will generally have lower reliability across occasions than it does for a single occasion.

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

In a computerized adaptive test such as STAR Reading, content varies from one administration to another, and it also varies with each student's performance. Another feature of computerized adaptive tests based on Item Response Theory

(IRT) is that the degree of measurement error can be expressed for each student's test individually.

The STAR Reading tests provide two ways to evaluate the reliability of scores: reliability coefficients, which indicate the overall precision of a set of test scores, and conditional standard errors of measurement (CSEM), which provide an index of the degree of error in an individual test score. A reliability coefficient is a summary statistic that reflects the average amount of measurement precision in a specific examinee group or in a population as a whole. In STAR Reading, the CSEM is an estimate of the unreliability of each individual test score. While a reliability coefficient is a single value that applies to the overall test, the magnitude of the CSEM may vary substantially from one person's test score to another's.

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

The reliability and measurement error presentation is divided into two sections below: First is a section describing the reliability coefficients and conditional standard errors of measurement for the original 25-item STAR Reading tests. Second, another brief section presents reliability and measurement error data for the newer, 34-item STAR Reading Enterprise tests.

25-Item STAR Reading Tests

Generic Reliability

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

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

where σ^2_{error} is the variance of the errors of measurement, and σ^2_{total} is the variance of test scores. In STAR Reading, the variance of the test scores is easily calculated from Scaled Score data. The variance of the errors of measurement may be estimated from the conditional standard error of measurement (CSEM) statistics that accompany each of the IRT-based test scores, including the Scaled Scores, as depicted below.

$$\sigma^2_{error} = \frac{1}{n}\sum CSEM_i^2$$

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

Using this technique with the STAR Reading 2008 norming data resulted in the generic reliability estimates shown in Table 17 on page 53. Because this method is not susceptible to error variance introduced by repeated testing, multiple occasions, and alternate forms, the resulting estimates of reliability are generally higher than the more conservative alternate forms reliability coefficients. These generic reliability coefficients are, therefore, plausible upper-bound estimates of the internal consistency reliability of the STAR Reading computer-adaptive test.

Generic reliability estimates are shown in Table 17. Results indicated that the overall reliability of the scores was about 0.95. Coefficients ranged from a low of 0.89 in grades 3 and 4 to a high of 0.93 in grades 10, 11, and 12. These reliability estimates are quite consistent across grades 1–12, and quite high for a test composed of only 25 items.

Overall, these coefficients also compare very favorably with the reliability estimates provided for other published reading tests, which typically contain far more items than the 25-item STAR Reading 4.3 and higher tests. The STAR Reading test’s high reliability with minimal testing time is a result of careful test item construction and an effective and efficient adaptive-branching procedure.

Split-Half Reliability

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

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

In internal simulation studies, the split-half method provided accurate estimates of the internal consistency reliability of adaptive tests, and so it has been used to provide estimates of STAR Reading reliability. These split-half reliability coefficients are independent of the generic reliability approach discussed earlier and more firmly grounded in the item response data. Split-half scores were based on the first 24 items of the STAR Reading norming test; scores based on the odd- and the even-numbered items were calculated separately. The correlations between the two sets of scores were corrected to a length of 25 items, yielding the split-half reliability estimates displayed in Table 17 on page 53.

Results indicated that the overall split-half reliability of the scores was about 0.92. The coefficients ranged from a low of 0.88 in grade 1 to a high of 0.91 in grade 12. These reliability estimates are quite consistent across grades 1–12, and quite high for a test composed of only 25 items, again a result of the measurement efficiency inherent in the adaptive nature of the STAR Reading test.

Alternate Form Reliability

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

Errors of measurement due to both content sampling and temporal changes in individuals' performance can affect alternate forms reliability coefficients, usually making them appreciably lower than internal consistency reliability coefficients. In addition, any growth in the trait that takes place in the interval between tests can also lower the correlation.

The alternate form reliability study provided estimates of STAR Reading 4.3 reliability using a variation of the test-retest method. In the traditional approach to test-retest reliability, students take the same test twice, with a short time interval, usually a few days, between administrations. In contrast, the STAR Reading 4.3 alternate form reliability study administered two different tests by avoiding during the second test the use of any items the student had encountered in the first test. All other aspects of the two tests were identical. The correlation coefficient between the scores on the two tests was taken as the reliability estimate.

The alternate form reliability estimates for the STAR Reading 4.3 test were calculated using the STAR Reading IRT ability estimates, or theta scores. Checks were made for valid test data on both test administrations and to remove cases of apparent motivational discrepancies.

Table 17 provides an overview of the reliability estimates for each grade along with an indication of the average number of days between testing occasions. The average number of days between testing occasions ranged from 5–8 days with most grades having taken the follow-up test about 1 week after the initial test. Results indicated that the overall reliability of the scores was about 0.91. The alternate form coefficients ranged from a low of 0.80 in grades 8, 10, and 11 to a high of 0.90 in grade 12.

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

Table 17: Reliability Estimates from the STAR Reading Norming Study: Spring 2008

		Reliability Estimates				
		Generic	Split-Half	Test-Retest		
Grade	N	ρxx	ρxx	N	ρxx	Average Days between Testing
1	7,523	0.91	0.88	298	0.89	8
2	10,132	0.9	0.89	296	0.85	7
3	10,476	0.89	0.89	297	0.82	7
4	9,984	0.89	0.89	297	0.83	7
5	8,352	0.9	0.89	300	0.83	7
6	6,462	0.9	0.89	294	0.81	7
7	4,767	0.91	0.9	288	0.83	7
8	4,364	0.91	0.9	284	0.8	7
9	2,921	0.92	0.9	241	0.86	8
10	2,079	0.93	0.9	214	0.8	7
11	1,795	0.93	0.9	209	0.8	5
12	1,153	0.93	0.91	245	0.9	8
Overall	69,738	0.95	0.92	3,263	0.91	7

Standard Error of Measurement

When interpreting the results of any test instrument, it is important to remember that the scores represent estimates of a student's true ability level. Test scores are not absolute or exact measures of performance. Nor is a single test score infallible in the information that it provides. The standard error of measurement can be thought of as a measure of how precise a given score is. The standard error of measurement describes the extent to which scores would be expected to fluctuate because of chance. If measurement errors follow a normal distribution, an SEM of 36 means that if a student were tested repeatedly, his or her scores would fluctuate within 36 points of his or her first score about 68 percent of the time, and within 72 points (twice the SEM) roughly 95 percent of the time. Since reliability can also be regarded as a measure of precision, there is a direct relationship between the reliability of a test and the standard error of measurement for the scores it produces.

The STAR Reading 4.3 and higher tests differ from traditional tests in at least two respects with regard to the standard error of measurement. First, STAR Reading software computes the SEM for each individual student based on his or her performance, unlike most printed tests that report the same SEM value for every examinee. Each administration of the test yields a unique “conditional” SEM that reflects the amount of information estimated to be in the specific combination of items that a student received in his or her individual test. Second, because the STAR Reading test is adaptive, the CSEM will tend to be lower than that of a conventional test, particularly at the highest and lowest score levels, where conventional tests’ measurement precision is weakest. Because the adaptive testing process attempts to provide equally precise measurement, regardless of the student’s ability level, the average CSEMs for the IRT ability estimates are very similar for all students. However, because the transformation of the STAR Reading IRT ability estimates into equivalent Scaled Scores is not linear, the SEMs in the Scaled Score metric are less similar.

Table 18 summarizes the distribution of CSEM values for the 2008 norms sample, overall and by grade level. The third through seventh columns provide the magnitude of the CSEM at the 5th, 25th, 50th (median), 75th, and 95th Percentile Ranks of the grade level distribution of conditional SEMs. The overall median CSEM across all grades was 51 scaled score units and ranged from a low of 36 in 1st grade to a high of 83 in 8th and 11th grades.

Table 18: Standard Error of Measurement by Selected Percentiles for Norming Sample Scaled Scores: Spring 2008

Grade	N	5th Percentile	25th Percentile	50th Percentile (Median)	75th Percentile	95th Percentile
1	7,523	5	28	36	40	52
2	10,132	20	36	41	49	62
3	10,476	34	40	48	57	84

Table 18: Standard Error of Measurement by Selected Percentiles for Norming Sample Scaled Scores: Spring 2008 (Continued)

Grade	N	5th Percentile	25th Percentile	50th Percentile (Median)	75th Percentile	95th Percentile
4	9,984	35	46	54	68	111
5	8,352	37	50	62	89	120
6	6,462	37	53	71	103	122
7	4,767	37	55	78	109	123
8	4,364	35	56	83	110	123
9	2,921	28	55	81	110	123
10	2,079	18	54	82	109	122
11	1,795	3	52	83	109	122
12	1,153	3	48	80	108	121
Overall	69,738	22	39	51	69	115

34-Item STAR Reading Enterprise Tests ENTERPRISE

Reliability Coefficients

STAR Reading Enterprise was designed to be a standards-based assessment, meaning that its item bank measures skills identified by exhaustive analysis of national and state standards in Reading, from grades K–12. STAR Reading Enterprise content covers many more skills than STAR Reading versions 1 through 4.4 RP.

The increased length of STAR Reading Enterprise, combined with its increased breadth of skills coverage and enhanced technical quality, was expected to result in improved measurement precision; this showed up as increased reliability, in both internal consistency reliability and alternate form reliability.

STAR Reading Enterprise test scores from tests administered in September 2012 through June 2013 were used to compute internal consistency and test-retest reliability estimates of STAR Reading Enterprise. Table 19 displays the estimated internal consistency and test-retest reliability of STAR Reading Enterprise test scores both overall and by grade.

Table 19: Reliability Estimates from the STAR Reading Enterprise Norming Study: Spring 2014

		Reliability Estimates			
		Generic	Test-Retest		
Grade	N	ρxx	N	ρxx	Average Days between Testing
1	100,000	0.95	8,000	0.8	92
2	100,000	0.94	8,000	0.85	97
3	100,000	0.93	8,000	0.85	98
4	100,000	0.93	8,000	0.85	99
5	100,000	0.93	8,000	0.86	99
6	100,000	0.93	8,000	0.87	104
7	100,000	0.93	8,000	0.87	107
8	100,000	0.94	8,000	0.87	106
9	100,000	0.94	8,000	0.87	114
10	100,000	0.94	8,000	0.87	116
11	100,000	0.95	8,000	0.86	117
12	100,000	0.95	8,000	0.85	112
Overall	1,200,000	0.97	96,000	0.93	105

As the table shows, STAR Reading Enterprise reliability is appreciably higher, grade by grade and overall, than the shorter 25-item STAR Reading version. STAR Reading Enterprise also demonstrates high test-retest consistency as shown in Table 19. The Enterprise version takes STAR Reading to new heights in technical quality, putting this interim assessment on a virtually equal footing with the highest-quality summative assessments in use today.

Standard Error of Measurement

Table 20 contains two different sets of estimates of STAR Reading Enterprise measurement error: conditional standard error of measurement (CSEM) and global standard error of measurement (SEM). Conditional SEM was described earlier in the introduction of this section on Reliability and Measurement Precision; the estimates of CSEM in Table 20 are the average CSEM values observed for each grade.

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

$$\text{SEM} = \text{SQRT}(1 - \rho)\ \sigma_x$$

where

SQRT() is the square root operator

ρ is the estimated internal consistency reliability

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

Global estimates of SEM can be expected to be more conservative (larger) than CSEM estimates, because the former are calculated from observed data, while the individual CSEM values are theory-based. To the extent that students' item responses do not perfectly fit the IRT model used (here, the Rasch model), CSEM may underestimate measurement error. Consistent with that, Table 20's global values of SEM are slightly greater than the counterpart CSEM values at every grade.

Comparing the estimates of reliability and measurement error of STAR Reading (Tables 17, 18) with those of STAR Reading Enterprise (Tables 19, 20) confirms that STAR Reading Enterprise is appreciably superior to the shorter STAR Reading assessments in terms of reliability and measurement precision.

Table 20: Estimates of STAR Reading Enterprise Measurement Precision by Grade and Overall: Conditional and Global Standard Error of Measurement

		Standard Error of Measurement		
		Conditional		
Grade	Sample Size	Average	Standard Deviation	Global
1	100,000	20	13.6	24
2	100,000	31	12.1	33
3	100,000	41	15.4	44
4	100,000	50	19.4	53
5	100,000	57	22.7	61
6	100,000	64	24.5	68
7	100,000	67	25.4	72
8	100,000	71	26.2	75
9	100,000	70	27	75
10	100,000	70	28.2	75
11	100,000	69	29	74
12	100,000	68	30.2	74
All	1,200,000	57	28.8	59

The National Center on Response to Intervention (NCRTI) and Screening

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

When evaluating screening tool reliability, NCRTI considered several factors:

- reliability of the performance level score
- disaggregated reliability data

NCRTI ratings include four qualitative labels: convincing evidence, partially convincing evidence, unconvincing evidence, or data unavailable/inadequate. Please refer to Table 21 for descriptions of these labels as provided by NCRTI, as well as the scores assigned to STAR Reading in each of the categories. Further descriptive information is provided within Tables 22 and 23.

Table 21: NCRTI Progress-Monitoring Indicator Descriptions

Indicator	Description	STAR Reading Score
Reliability of the Performance Level Score	Reliability of the performance level score is the extent to which the score (or average/median of 2–3 scores) is accurate and consistent.	Convincing Evidence
Disaggregated Reliability Data	Disaggregated data are scores that are calculated and reported separately for specific subgroups (e.g., race, economic status, special education status, etc.).	Convincing Evidence

Table 22: Reliability of the Performance Level Score for STAR Reading

Type of Reliability	Grade	N (Range)	Coefficient		SEM	Information (Including Normative Data)/Subjects
			Range	Median		
Generic	1–5	7,523–10,476	0.89–0.91	0.90	36–62 Median: 48	Based on STAR Reading 4.3 norms sample, IRT reliability was calculated from the conditional error variance of IRT ability estimates.
Split-Half	1–5	7,523–10,476	0.88–0.89	0.89	NA	Split-half reliability was calculated in the 2.0 norming sample.
Retest	1–5	296–300	0.82–0.89	0.83	NA	There were no common items across retests; non-overlapping versions of STAR Reading 4.3 were taken.

Table 22: Reliability of the Performance Level Score for STAR Reading (Continued)

Type of Reliability	Grade	N (Range)	Coefficient		SEM	Information (Including Normative Data)/Subjects
			Range	Median		
Generic	6–12	1,153–6,462	0.90–0.93	0.92	71–83 Median: 81	Based on STAR Reading 4.3 norms sample, IRT reliability was calculated from the conditional error variance of IRT ability estimates.
Split-Half	6–12	1,153–6,462	0.89–0.91	0.90	NA	Split-half reliability was calculated in the 2.0 norming sample.
Retest	6–12	209–294	0.80–0.90	0.81	NA	There were no common items across retests; non-overlapping versions of STAR Reading 4.3 were taken.

Table 23: Disaggregated Reliability of the Performance Level Score for STAR Reading

Type of Reliability	Grade	N (Range)	Coefficient		SEM	Information (Including Normative Data)/Subjects
			Range	Median		
Generic (White)	1–5	114,297	0.87–0.91	0.87	50	Data from spring 2008 STAR Reading assessments of 1,864 different customers representing 50 states and Canada. Of this sample 21% were Black, 30% Hispanic, and 49% white.
Generic (Black)		48,718	0.89–0.89	0.89	42	
Generic (Hispanic)		67,456	0.89–0.90	0.89	41	
Generic (White)	6–12	36,915	0.88–0.94	0.90	90	
Generic (Black)		15,632	0.90–0.94	0.91	65	
Generic (Hispanic)		24,628	0.91–0.94	0.92	61	

The National Center on Intensive Intervention (NCII) and Progress Monitoring

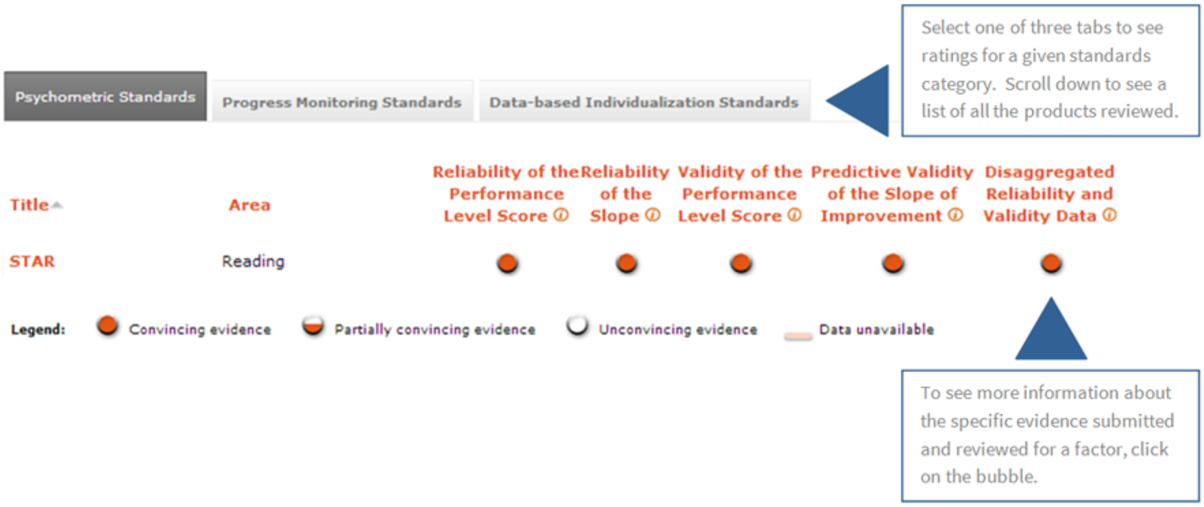
NCII is a more recent federally funded project; it is related to NCRTI but was created in 2012 with a mission focusing on just those students with severe learning needs. NCII reviews are currently ongoing and focus on the technical adequacy of assessments as progress-monitoring tools. The technical criteria and rating system were carried over from NCRTI and STAR Reading has again earned strong ratings.

When evaluating progress monitoring tools, NCII considers a variety of factors in three general standards categories:

- Psychometric Standards
- Progress Monitoring Standards
- Data-Based Individualization Standards

Please refer to the NCII website for the most up to date information about the factors included in reviews and scores assigned to STAR Reading: http://www.intensiveintervention.org/chart/progress-monitoring. Figure 2 provides a snapshot of the NCII website navigation features.

Figure 2: Screenshot from NCII Website: Academic Progress Monitoring General Outcome Measures



Psychometric Standards
Progress Monitoring Standards
Data-based Individualization Standards

Title▲	Area	Reliability of the Performance Level Score ⓘ	Reliability of the Slope ⓘ	Validity of the Performance Level Score ⓘ	Predictive Validity of the Slope of Improvement ⓘ	Disaggregated Reliability and Validity Data ⓘ
STAR	Reading	●	●	●	●	●

Legend: ● Convincing evidence ◒ Partially convincing evidence ○ Unconvincing evidence ▭ Data unavailable

Select one of three tabs to see ratings for a given standards category. Scroll down to see a list of all the products reviewed.

To see more information about the specific evidence submitted and reviewed for a factor, click on the bubble.

Validity

The key concept often used to judge an instrument's usefulness is its validity. The validity of a test is the degree to which it assesses what it claims to measure. Determining the validity of a test involves the use of data and other information both internal and external to the test instrument itself. One touchstone is content validity, which is the relevance of the test questions to the attributes supposed to be measured by the test—namely reading comprehension and reading achievement, in the case of the STAR Reading test. These content validity issues were discussed in detail in "Content and Item Development" (beginning on page 19) and were an integral part of the test items that form the basis of STAR Reading versions 2.0 through 4.4, as well as the new STAR Reading Enterprise version.

Construct validity, which is the overarching criterion for evaluating a test, investigates the extent to which a test measures the construct that it claims to be assessing. Establishing construct validity involves the use of data and other information external to the test instrument itself. For example, STAR Reading versions 2.0 through 4.4 claim to provide an estimate of a child's reading comprehension and achievement level. Therefore, demonstration of STAR Reading's construct validity rests on the evidence that the test provides such estimates. There are a number of ways to demonstrate this.

For instance, in a study linking STAR Reading and the Degrees of Reading Power comprehension assessment, a raw correlation of 0.89 was observed between the two tests. Adjusting that correlation for attenuation due to unreliability yielded a corrected correlation of 0.96, indicating that the constructs (i.e., reading comprehension) measured by STAR Reading and Degrees of Reading Power are almost indistinguishable. Table 26 on page 71 and Table 27 on page 75 present evidence of predictive validity collected subsequent to the SR 2.0 norming study. These two tables display numerous correlations between STAR Reading and other measures administered at points in time at least two months later than STAR Reading.

Since reading ability varies significantly within and across grade levels and improves as a student's grade placement increases, scores within STAR Reading should demonstrate these anticipated internal relationships; in fact, they do. Additionally, scores within STAR Reading should correlate highly with other accepted procedures and measures that are used to determine reading achievement and reading comprehension; this is external validity.

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

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

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

In an ongoing effort to gather evidence for the validity of STAR Reading scores, continual research on score validity has been undertaken. In addition to original validity data gathered at the time of initial development, numerous other studies have investigated the correlations between STAR Reading tests and other external measures. In addition to gathering concurrent validity estimates, predictive validity estimates have also been investigated. Concurrent validity was defined for students taking a STAR Reading test and external measures within a two-month time period. Predictive validity provided an estimate of the extent to which scores on the STAR Reading test predicted scores on criterion measures given at a later point in time, operationally defined as more than two months between the STAR test (predictor) and the criterion test. 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 24–27 present the correlation coefficients between the scores on the STAR Reading 2.0 test and each of the other tests for which data were received. Tables 24 and 25 display "concurrent validity" data; that is, correlations between STAR Reading test scores and other tests administered within a two-month time period. The date of administration ranged from spring 1999–spring 2013. More recently, data have become available for analyses regarding the predictive validity of STAR Reading. Predictive validity provides an estimate of the extent to which scores on the STAR Reading test predicted scores on criterion measures given at a later point in time, operationally defined as more than 2 months between the STAR test (predictor) and the criterion test. Predictive validity provides an estimate of the linear relationship between STAR scores and scores on tests covering a similar academic domain. Predictive correlations are attenuated by time due to the fact that students are gaining skills in the interim between testing occasions, and also by differences between the tests' content specifications. Tables 26 and 27 present predictive validity coefficients.

Tables 24–27 are presented in two parts. Tables 24 and 26 display validity coefficients for grades 1–6, and Tables 25 and 27 display the validity coefficients for grades 7–12. The bottom of each table presents a grade-by-grade summary, including the total number of students for whom test data were available, the number of validity coefficients for that grade, and the average value of the validity coefficients.

The within-grade average concurrent validity coefficients for grades 1–6 varied from 0.72–0.80, with an overall average of 0.74. The within-grade average concurrent validity for grades 7–12 ranged from 0.65–0.76, with an overall average of 0.72. Predictive validity coefficients ranged from 0.69–0.72 in grades 1–6, with an average of 0.71. In grades 7–12 the predictive validity coefficients ranged from 0.72–0.87 with an average of 0.80. The other validity coefficient within-grade averages (for STAR Reading 2.0 with external tests administered prior to spring 1999, Tables 28 and 29) varied from 0.60–0.77; the overall average was 0.72. The process of establishing the validity of a test is laborious, and it usually takes a significant amount of time. As a result, the validation of the STAR Reading test is an ongoing activity, with the goal of establishing evidence of the test's validity for a variety of settings and students. STAR Reading users who collect relevant data are encouraged to contact Renaissance Learning.

Since correlation coefficients are available for many different test editions, forms, and dates of administration, many of the tests have several validity coefficients associated with them. Data were omitted from the tabulations if (a) test data quality could not be verified or (b) when sample size was very small. Testing data for other standardized tests administered prior to spring 2006 were excluded from the validity analyses. In general, these correlation coefficients reflect very well on the validity of the STAR Reading test as a tool for placement in Reading. In fact, the correlations are similar in magnitude to the validity coefficients of these measures with each other. These validity results, combined with the supporting evidence of reliability and minimization of SEM estimates for the STAR Reading test, provide a quantitative demonstration of how well this innovative instrument in reading achievement assessment performs.

Table 24: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a]

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
Arkansas Augmented Benchmark Examination (AABE)														
AABE	S 08	SS	–	–	–	–	2,858	0.78*	2,588	0.73*	1,897	0.73*	1,176	0.75*
AIMSweb														
R-CBM	S 12	correct	15	0.65*	72	0.28*	41	0.17	44	0.48*	–	–	–	–
California Achievement Test (CAT)														
CAT	S 99	SS	93	0.80*	36	0.67*	–	–	34	0.72*	146	0.76*	–	–
CAT/5	F 10–11	SS	68	0.79*	315	0.72*	410	0.69*	302	0.71*	258	0.71*	196	0.69*
Canadian Achievement Test (CAT)														
CAT/2	F 10–11		–	–	–	–	21	0.80*	31	0.84*	23	0.75*	–	–
Colorado Student Assessment Program (CSAP)														
CSAP	S 06	SS	–	–	–	–	82	0.75*	79	0.83*	93	0.68*	280	0.80*
Comprehensive Test of Basic Skills (CTBS)														
CTBS/4	S 99	NCE	–	–	–	–	–	–	18	0.81*	–	–	–	–
CTBS/A-19/20	S 99	SS	–	–	–	–	–	–	–	–	–	–	8	0.91*
Delaware Student Testing Program (DSTP) – Reading														
DSTP	S 05	SS	–	–	–	–	104	0.57*	–	–	–	–	–	–
DSTP	S 06	SS	–	–	158	0.68*	126	0.43*	141	0.62*	157	0.59*	75	0.66*
Dynamic Indicators of Basic Early Literacy Skills (DIBELS) – Oral Reading Fluency														
DIBELS	F 05	WCPM	–	–	59	0.78*	–	–	–	–	–	–	–	–
DIBELS	W 06	WCPM	61	0.87*	55	0.75*	–	–	–	–	–	–	–	–
DIBELS	S 06	WCPM	67	0.87*	63	0.71*	–	–	–	–	–	–	–	–
DIBELS	F 06	WCPM	–	–	515	0.78*	354	0.81*	202	0.72*	–	–	–	–
DIBELS	W 07	WCPM	208	0.75*	415	0.73*	175	0.69*	115	0.71*	–	–	–	–
DIBELS	S 07	WCPM	437	0.81*	528	0.70*	363	0.66*	208	0.54*	–	–	–	–
DIBELS	F 07	WCPM	–	–	626	0.79*	828	0.73*	503	0.73*	46	0.73*	–	–

Table 24: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a] (Continued)

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
Florida Comprehensive Assessment Test (FCAT)														
FCAT	S 06	SS	–	–	–	–	–	–	41	0.65*	–	–	–	–
FCAT	S 06–08	SS	–	–	–	–	10,169	0.76*	8,003	0.73*	5,474	0.73*	1,188	0.67*
Florida Comprehensive Assessment Test (FCAT 2.0)														
FCAT 2.0	S 13	SS	–	–	–	–	3,641	0.83*	3,025	0.84*	2,439	0.83*	145	0.81*
Gates–MacGinitie Reading Test (GMRT)														
GMRT/2nd Ed	S 99	NCE	–	–	21	0.89*	–	–	–	–	–	–	–	–
GMRT/L-3rd	S 99	NCE	–	–	127	0.80*	–	–	–	–	–	–	–	–
Idaho Standards Achievement Test (ISAT)														
ISAT	S 07–09	SS	–	–	–	–	3,724	0.75*	2,956	0.74*	2,485	0.74*	1,309	0.75*
Illinois Standards Achievement Test – Reading														
ISAT	S 05	SS	–	–	106	0.71*	594	0.76*	–	–	449	0.70*	–	–
ISAT	S 06	SS	–	–	–	–	140	0.80*	144	0.80*	146	0.72	–	–
Iowa Test of Basic Skills (ITBS)														
ITBS–Form K	S 99	NCE	40	0.75*	36	0.84*	26	0.82*	28	0.89*	79	0.74*	–	–
ITBS–Form L	S 99	NCE	–	–	–	–	18	0.70*	29	0.83*	41	0.78*	38	0.82*
ITBS–Form M	S 99	NCE	–	–	–	–	158	0.81*	–	–	125	0.84*	–	–
ITBS–Form K	S 99	SS	–	–	58	0.74*	–	–	54	0.79*	–	–	–	–
ITBS–Form L	S 99	SS	–	–	–	–	45	0.73*	–	–	–	–	50	0.82*
Kansas State Assessment Program (KSAP)														
KSAP	S 06–08	SS	–	–	–	–	4,834	0.61*	4,045	0.61*	3,332	0.63*	1,888	0.65*
Kentucky Core Content Test (KCCT)														
KCCT	S 08–10	SS	–	–	–	–	10,776	0.60*	8,885	0.56*	7,147	0.53*	5,003	0.57*
Metropolitan Achievement Test (MAT)														
MAT–7th Ed.	S 99	NCE	–	–	–	–	–	–	46	0.79*	–	–	–	–
MAT–6th Ed.	S 99	Raw	–	–	–	–	8	0.58*	–	–	8	0.85*	–	–
MAT 7th Ed.	S 99	SS	–	–	–	–	25	0.73*	17	0.76*	21	0.76*	23	0.58*

Table 24: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a] (Continued)

			1		2		3		4		5		6	
Test Form	**Date**	**Score**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**
Michigan Educational Assessment Program (MEAP) – English Language Arts														
MEAP	F 04	SS	–	–	–	–	–	–	155	0.81*	–	–	–	–
MEAP	F 05	SS	–	–	–	–	218	0.76*	196	0.80*	202	0.80*	207	0.69*
MEAP	F 06	SS	–	–	–	–	116	0.79*	132	0.69*	154	0.81*	129	0.66*
Michigan Educational Assessment Program (MEAP) – Reading														
MEAP	F 04	SS	–	–	–	–	–	–	155	0.80*	–	–	–	–
MEAP	F 05	SS	–	–	–	–	218	0.77*	196	0.78*	202	0.81*	207	0.68*
MEAP	F 06	SS	–	–	–	–	116	0.75*	132	0.70*	154	0.82*	129	0.70*
Mississippi Curriculum Test (MCT2)														
MCT2	S 02	SS	–	–	–	–	–	–	155	0.80*	–	–	–	–
MCT2	S 03	SS	–	–	–	–	218	0.77*	196	0.78*	202	0.81*	207	0.68*
MCT2	S 08	SS	–	–	–	–	3,821	0.74*	3,472	0.73*	2,915	0.71*	2367	0.68*
Missouri Mastery Achievement Test (MMAT)														
MMAT	S 99	NCE	–	–	–	–	–	–	–	–	26	0.62*	–	–
New Jersey Assessment of Skills and Knowledge (NJ ASK)														
NJ ASK	S 13	SS	–	–	–	–	1,636	0.79*	1,739	0.80*	1,486	0.82*	440	0.77*
New York State Assessment Program														
NYSTP	S 13	SS	–	–	–	–	185	0.78*	–	–	–	–	–	–
North Carolina End–of–Grade (NCEOG): Test														
	S 99	SS	–	–	–	–	–	–	–	–	85	0.79*	–	–
NCEOG	S 06–08	SS	–	–	–	–	2,707	0.80*	2,234	0.77*	1,752	0.77*	702	0.77*
Ohio Achievement Assessment (OAA)														
OAA	S 13	SS	–	–	–	–	1,718	0.72*	1,595	0.71*	1,609	0.77*	1,599	0.76*
Oklahoma Core Curriculum Test (OCCT)														
OCCT	S 06	SS	–	–	–	–	78	0.62*	92	0.58*	46	0.52*	80	0.60*
OCCT	S 13	SS	–	–	–	–	153	0.79*	66	0.79*	72	0.80*	64	0.72*

Table 24: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a] (Continued)

			1		2		3		4		5		6	
Test Form	**Date**	**Score**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**
South Dakota State Test of Educational Progress (DSTEP)														
DSTEP	S 08–10	SS	–	–	–	–	2,072	0.78*	1,751	0.77*	1,409	0.80*	906	0.78*
Stanford Achievement Test (SAT)														
SAT 9th Ed.	S 99	NCE	68	0.79*	–	–	26	0.44*	–	–	–	–	86	0.65*
SAT 9th Ed.	S 99	SS	11	0.89*	18	0.89*	67	0.79*	66	0.79*	72	0.80*	64	0.72*
State of Texas Assessments of Academic Readiness Standards Test (STAAR)														
STAAR	S 12–13	SS	–	–	–	–	8,567	0.79*	7,902	0.78*	7,272	0.76*	5,697	0.78*
Tennessee Comprehensive Assessment Program (TCAP)														
TCAP	S 11	SS	–	–	–	–	62	0.66*	56	0.59*	–	–	–	–
TCAP	S 12	SS	–	–	–	–	91	0.79*	118	0.21*	81	0.64*	–	–
TCAP	S 13	SS	–	–	–	–	494	0.73*	441	0.66*	426	0.77*	–	–
TerraNova														
TerraNova	S 99	SS	–	–	61	0.72*	117	0.78*	–	–	–	–	–	–
Texas Assessment of Academic Skills (TAAS)														
TAAS	S 99	NCE	–	–	–	–	–	–	–	–	–	–	229	0.66*
Transitional Colorado Assessment Program (TCAP)														
TCAP	S 12–13	SS	–	–	–	–	3,144	0.78*	3,200	0.82*	3,186	0.81*	3,106	0.83*
West Virginia Educational Standards Test 2 (WESTEST 2)														
WESTEST 2	S 12	SS	–	–	–	–	2,949	0.76*	7,537	0.77*	5,666	0.76*	2,390	0.75*
Woodcock Reading Mastery (WRM)														
	S 99		–	–	–	–	–	–	–	–	7	0.68*	7	0.66*
Wisconsin Knowledge and Concepts Examination (WKCE)														
WKCE	F 06–10	SS					8,649	0.78*	7,537	0.77*	5,666	0.76*	2,390	0.75*

Table 24: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 1–6[a] (Continued)

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

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

Table 25: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 7–12[a]

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

Table 25: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 7–12[a] (Continued)

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

Table 25: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 7–12[a] (Continued)

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Mississippi Curriculum Test (MCT2)														
MCT2	S 03	SS	372	0.70*	–	–	–	–	–	–	–	–	–	–
MCT2	S 08	SS	1,424	0.69*	1,108	0.72*	–	–	–	–	–	–	–	–
Missouri Mastery Achievement Test (MMAT)														
MMAT	S 99	NCE	–	–	29	0.78*	19	0.71*	–	–	–	–	–	–
North Carolina End-of-Grade (NCEOG) Test														
NCEOG	S 06–08	SS	440	0.76*	493	0.74*	–	–	–	–	–	–	–	–
New Jersey Assessment of Skills and Knowledge (NJ ASK)														
NJ ASK	S 13	SS	595	0.78*	589	0.70*	–	–	–	–	–	–	–	–
Northwest Evaluation Association Levels Test (NWEA)														
NWEA-Achieve	S 99	NCE	–	–	124	0.66*	–	–	–	–	–	–	–	–
South Dakota State Test of Educational Progress (DSTEP)														
DSTEP	S 08–10	SS	917	0.78*	780	0.77*	–	–	–	–	–	–	–	–
Stanford Achievement Test (SAT)														
SAT–9th Ed.	S 99	NCE	50	0.65*	50	0.51*	–	–	–	–	–	–	–	–
SAT–9th Ed.	S 99	SS	70	0.70*	68	0.80*	–	–	–	–	–	–	–	–
State of Texas Assessments of Academic Readiness Standards Test (STAAR)														
STAAR	S 12–13	SS	5,062	.075*	4,651	0.75*	–	–	–	–	–	–	–	–
Test Achievement and Proficiency (TAP)														
TAP	S 99	NCE	–	–	–	–	6	0.42	13	0.80*	7	0.6	–	–
Texas Assessment of Academic Skills (TAAS)														
TAAS	S 99	NCE	–	–	–	–	–	–	43	0.60*	–	–	–	–
Transitional Colorado Assessment Program (TCAP)														
TCAP	S 12–13	SS	3,165	0.83*	3,106	0.83*	1,466	0.72*	–	–	–	–	–	–
West Virginia Educational Standards Test 2 (WESTEST 2)														
WESTEST 2	S 12	SS	1,612	0.76	1,396	0.75	–	–	–	–	–	–	–	–

Table 25: Concurrent Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Spring 1999–Spring 2013, Grades 7–12[a] (Continued)

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Wisconsin Knowledge and Concepts Examination (WKCE)														
WKCE	F 06–10	SS	1,811	0.81	1,886	0.77	–	–	506	0.79	–	–	–	–
Wide Range Achievement Test 3 (WRAT3)														
WRAT3	S 99		–	–	17	0.81*	–	–	–	–	–	–	–	–
Summary														
Grade(s)		**All**	**7**		**8**		**9**		**10**		**11**		**12**	
Number of students		48,789	25,032		21,134		1,774		755		55		39	
Number of coefficients		74	30		29		7		5		2		1	
Average validity		–	0.74		0.73		0.65		0.76		0.70		0.73	
Overall average		0.72												

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

Table 26: Predictive Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 1–6[a]

Test Form	Date[b]	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
AIMSweb														
R-CBM	S 12	correct	60	0.14	156	0.38*	105	0.11	102	0.52*	–	–	–	–
Arkansas Augmented Benchmark Examination (AABE)														
AABE	F 07	SS	–	–	–	–	5,255	0.79*	5,208	0.77*	3,884	0.75*	3,312	0.75*
Colorado Student Assessment Program (CSAP)														
CSAP	F 04	–	–	–	–	–	82	0.72*	79	0.77*	93	0.70*	280	0.77*
Delaware Student Testing Program (DSTP) – Reading														
DSTP	S 05	–	–	–	–	–	189	0.58*	–	–	–	–	–	–
DSTP	W 05	–	–	–	–	–	120	0.67*	–	–	–	–	–	–
DSTP	S 05	–	–	–	–	–	161	0.52*	191	0.55*	190	0.62*	–	–
DSTP	F 05	–	–	–	253	0.64*	214	0.39*	256	0.62*	270	0.59*	242	0.71*
DSTP	W 05	–	–	–	275	0.61*	233	0.47*	276	0.59*	281	0.62*	146	0.57*

Table 26: Predictive Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 1–6[a] (Continued)

Test Form	Date[b]	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
Florida Comprehensive Assessment Test (FCAT)														
FCAT	F 05	–	–	–	–	–	–	–	42	0.73*	–	–	409	0.67*
FCAT	W 07	–	–	–	–	–	–	–	–	–	–	–	417	0.76*
FCAT	F 05–07	SS	–	–	–	–	25,192	0.78*	21,650	0.75*	17,469	0.75*	9,998	0.73*
Florida Comprehensive Assessment Test (FCAT 2.0)														
FCAT 2.0	S 13	SS	–	–	–	–	6,788	0.78*	5,894	0.80*	5,374	0.80*	616	0.74*
Idaho Standards Achievement Test (ISAT)														
ISAT	F 08–10	SS	–	–	–	–	8,219	0.77*	8,274	0.77*	7,537	0.76*	5,742	0.77*
Illinois Standards Achievement Test (ISAT) – Reading														
ISAT–R	F 05	–	–	–	–	–	450	0.73*	–	–	317	0.68*	–	–
ISAT–R	W 05	–	–	–	–	–	564	0.76*	–	–	403	0.68*	–	–
ISAT–R	F 05	–	–	–	–	–	133	0.73*	140	0.74*	145	0.66*	–	–
ISAT–R	W 06	–	–	–	–	–	138	0.76*	145	0.77*	146	0.70*	–	–
Iowa Assessment														
IA	F 12	SS	–	–	–	–	1,763	0.61*	1,826	0.61*	1,926	0.59*	1,554	0.64*
IA	W 12	SS	–	–	–	–	548	0.60*	661	0.62*	493	0.64*	428	0.65*
IA	S 12	SS	–	–	–	–	1,808	0.63*	1,900	0.63*	1,842	0.65*	1,610	0.63*
Kentucky Core Content Test (KCCT)														
KCCT	F 07–09	SS	–	–	–	–	16,521	0.62*	15,143	0.57*	12,549	0.53*	9,091	0.58*
Michigan Educational Assessment Program (MEAP) – English Language Arts														
MEAP–EL	F 04	–	–	–	–	–	193	0.60*	181	0.70*	170	0.75*	192	0.66*
MEAP–EL	W 05	–	–	–	–	–	204	0.68*	184	0.74*	193	0.75*	200	0.70*
MEAP–EL	S 05	–	–	–	–	–	192	0.73*	171	0.73*	191	0.71*	193	0.62*
MEAP–EL	F 05	–	–	–	–	–	111	0.66*	132	0.71*	119	0.77*	108	0.60*
MEAP–EL	W 06	–	–	–	–	–	114	0.77*	–	–	121	0.75*	109	0.66*

Table 26: Predictive Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 1–6[a] (Continued)

Test Form	Date[b]	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
Michigan Educational Assessment Program (MEAP) – Reading														
MEAP–R	F 04	–	–	–	–	–	193	0.60*	181	0.69*	170	0.76*	192	0.66*
MEAP–R	W 05	–	–	–	–	–	204	0.69*	184	0.74*	193	0.78*	200	0.70*
MEAP–R	S 05	–	–	–	–	–	192	0.72*	171	0.72*	191	0.74*	193	0.62*
MEAP–R	F 05	–	–	–	–	–	111	0.63*	132	0.70*	119	0.78*	108	0.62*
MEAP–R	W 06	–	–	–	–	–	114	0.72*	–	–	121	0.75*	109	0.64*
Mississippi Curriculum Test (MCT2)														
MCT2	F 01	–	–	–	86	0.57*	95	0.70*	97	0.65*	78	0.76*	–	–
MCT2	F 02	–	–	–	340	0.67*	337	0.67*	282	0.69*	407	0.71*	442	0.72*
MCT2	F 07	SS	–	–	–	–	6,184	0.77*	5,515	.74*	5,409	0.74*	4,426	0.68*
North Carolina End–of–Grade (NCEOG) Test														
NCEOG	F 05–07	SS	–	–	–	–	6,976	0.81*	6,531	0.78*	6,077	0.77*	3,255	0.77*
New York State Assessment Program														
NYSTP	S 13	SS	–	–	–	–	349	0.73*	–	–	–	–	–	–
Ohio Achievement Assessment														
OAA	S 13	SS	–	–	–	–	28	0.78*	41	0.52*	29	0.79*	30	0.75*
Oklahoma Core Curriculum Test (OCCT)														
OCCT	F 04	–	–	–	–	–	–	–	–	–	44	0.63*	–	–
OCCT	W 05	–	–	–	–	–	–	–	–	–	45	0.66*	–	–
OCCT	F 05	–	–	–	–	–	89	0.59*	90	0.60*	79	0.69*	84	0.63*
OCCT	W 06	–	–	–	–	–	60	0.65*	40	0.67*	–	–	–	–
South Dakota State Test of Educational Progress (DSTEP)														
DSTEP	F 07–09	SS	–	–	–	–	3,909	0.79*	3,679	0.78*	3,293	0.78*	2,797	0.79*

Table 26: Predictive Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 1–6[a] (Continued)

Test Form	Date[b]	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
STAR Reading														
STAR–R	F 05	–	16,982	0.66*	42,601	0.78*	46,237	0.81*	44,125	0.83*	34,380	0.83*	23,378	0.84*
STAR–R	F 06	–	25,513	0.67*	63,835	0.78*	69,835	0.81*	65,157	0.82*	57,079	0.83*	35,103	0.83*
STAR–R	F 05	–	8,098	0.65*	20,261	0.79*	20,091	0.81*	18,318	0.82*	7,621	0.82*	5,021	0.82*
STAR–R	F 05	–	8,098	0.55*	20,261	0.72*	20,091	0.77*	18,318	0.80*	7,621	0.80*	5,021	0.79*
STAR–R	S 06	–	8,098	0.84*	20,261	0.82*	20,091	0.83*	18,318	0.83*	7,621	0.83*	5,021	0.83*
STAR–R	S 06	–	8,098	0.79*	20,261	0.80*	20,091	0.81*	18,318	0.82*	7,621	0.82*	5,021	0.81*
State of Texas Assessments of Academic Readiness Standards Test (STAAR)														
STAAR	S 12–13	SS	–	–	–	–	6,132	0.81*	5,744	0.80*	5,327	0.79*	5,143	0.79*
Tennessee Comprehensive Assessment Program (TCAP)														
TCAP	S 11	SS	–	–	–	–	695	0.68*	602	0.72*	315	0.61*	–	–
TCAP	S 12	SS	–	–	–	–	763	0.70*	831	0.33*	698	0.65*	–	–
TCAP	S 13	SS	–	–	–	–	2,509	0.67*	1,897	0.63*	1,939	0.68*	431	0.65*
West Virginia Educational Standards Test 2 (WESTEST 2)														
WESTEST 2	S 12	SS	–	–	–	–	2,828	0.80*	3,078	0.73*	3,246	0.73*	3,214	0.73*
Wisconsin Knowledge and Concepts Examination (WKCE)														
WKCE	S 05–09	SS					15,706	0.75*	15,569	0.77*	13,980	0.78*	10,641	0.78*
Summary														
Grade(s)	**All**		**1**		**2**		**3**		**4**		**5**		**6**	
Number of students	1,227,887		74,887		188,434		313,102		289,571		217,416		144,477	
Number of coefficients	194		6		10		49		43		47		39	
Average validity			0.69		0.72		0.70		0.71		0.72		0.71	
Overall average	0.71													

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

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

Table 27: Predictive Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 7–12[a]

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

Table 27: Predictive Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 7–12[a] (Continued)

Test Form	Date[b]	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Michigan Educational Assessment Program (MEAP) – English Language Arts														
MEAP	F 04	–	181	0.71*	88	0.85*	–	–	–	–	–	–	–	–
MEAP	W 05	–	214	0.73*	212	0.73*	–	–	–	–	–	–	–	–
MEAP	S 05	–	206	0.75*	223	0.69*	–	–	–	–	–	–	–	–
MEAP	F 05	–	114	0.66*	126	0.66*	–	–	–	–	–	–	–	–
MEAP	W 06	–	114	0.64*	136	0.71*	–	–	–	–	–	–	–	–
MEAP	S 06	–	–	–	30	0.80*	–	–	–	–	–	–	–	–
Michigan Educational Assessment Program (MEAP) – Reading														
MEAP–R	F 04	–	181	0.70*	88	0.84*	–	–	–	–	–	–	–	–
MEAP–R	W 05	–	214	0.72*	212	0.73*	–	–	–	–	–	–	–	–
MEAP–R	S 05	–	206	0.72*	223	0.69*	–	–	–	–	–	–	–	–
MEAP–R	F 05	–	116	0.68*	138	0.66*	–	–	–	–	–	–	–	–
MEAP–R	W 06	–	116	0.68*	138	0.70*	–	–	–	–	–	–	–	–
MEAP–R	S 06	–	–	–	30	0.81*	–	–	–	–	–	–	–	–
Mississippi Curriculum Test (MCT2)														
MCT2	F 02	–	425	0.68*	–	–	–	–	–	–	–	–	–	–
MCT2	F 07	SS	3,704	0.68*	3,491	0.73*	–	–	–	–	–	–	–	–
North Carolina End–of–Grade (NCEOG) Test														
NCEOG	F 05–07	SS	2,735	0.77*	2,817	0.77*	–	–	–	–	–	–	–	–
Ohio Achievement Assessment														
OAA	S 13	SS	53	0.82*	38	0.66*	–	–	–	–	–	–	–	–
South Dakota State Test of Educational Progress (DSTEP)														
DSTEP	F 07–09	SS	2,236	0.79*	2,073	0.78*	–	–	–	–	–	–	–	–

Table 27: Predictive Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Fall 2005–Spring 2013, Grades 7–12[a] (Continued)

Test Form	Date[b]	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
STAR Reading														
STAR–R	F 05	–	17,370	0.82*	9,862	0.82*	2,462	0.82*	15,277	0.85*	1,443	0.83*	596	0.85*
STAR–R	F 06	–	22,177	0.82*	19,152	0.82*	4,087	0.84*	2,624	0.85*	2,930	0.85*	2,511	0.86*
STAR–R	F 05	–	5,399	0.81*	641	0.76*	659	0.89*	645	0.88*	570	0.90*	–	–
STAR–R	F 05	–	5,399	0.79*	641	0.76*	659	0.83*	645	0.83*	570	0.87*	–	–
STAR–R	S 06	–	5,399	0.82*	641	0.83*	659	0.87*	645	0.88*	570	0.89*	–	–
STAR–R	S 06	–	5,399	0.80*	641	0.83*	659	0.85*	645	0.85*	570	0.86*		
State of Texas Assessments of Academic Readiness Standards Test (STAAR)														
STAAR	S 12–13	SS	4,716	0.77*	4,507	0.76*	–	–	–	–	–	–	–	–
Tennessee Comprehensive Assessment Program (TCAP)														
TCAP	S 13	SS	332	0.81*	233	0.74*	–	–	–	–	–	–	–	–
West Virginia Educational Standards Test 2 (WESTEST 2)														
WESTEST 2	S 12	SS	2,852	0.71*	2,636	0.74*	–	–	–	–	–	–	–	–
Wisconsin Knowledge and Concepts Examination (WKCE)														
WKCE	S 05–09	SS	6,399	0.78*	5,500	0.78*			401	0.78*				
Summary														
Grade(s)	**All**		**7**		**8**		**9**		**10**		**11**		**12**	
Number of students	224,179		111,143		72,537		9,567		21,172		6,653		3,107	
Number of coefficients	106		39		41		8		10		6		2	
Average validity	–		0.72		0.73		0.81		0.81		0.87		0.86	
Overall average	0.80													

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

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

Table 28: Other External Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6[a]

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

Table 28: Other External Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6[a] (Continued)

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

Table 28: Other External Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6[a] (Continued)

			1		2		3		4		5		6	
Test Form	**Date**	**Score**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**	**n**	**r**
TerraNova														
	Fall 97	Scaled	–	–	–	–	–	–	–	–	56	0.70*	–	–
	Spr 98	NCE	–	–	–	–	76	0.63*	–	–	–	–	–	–
	Spr 98	Scaled	–	–	94	0.50*	55	0.79*	299	0.75*	86	0.75*	23	0.59*
	Fall 98	NCE	–	–	–	–	–	–	–	–	–	–	126	0.74*
	Fall 98	Scaled	–	–	–	–	–	–	14	0.70*	–	–	15	0.77*
Wide Range Achievement Test 3 (WRAT3)														
	Fall 98		–	–	–	–	–	–	–	–	–	–	10	0.89*
Wisconsin Reading Comprehension Test														
	Spr 98		–	–	–	–	–	–	63	0.58*	–	–	–	–
Summary														
Grade(s)		**All**	**1**		**2**		**3**		**4**		**5**		**6**	
Number of students		4,289	150		691		734		1,091		871		752	
Number of coefficients		95	7		14		19		16		18		21	
Average validity		–	0.75		0.72		0.73		0.74		0.73		0.71	
Overall average		0.73												

a. Sample sizes are in the columns labeled "n."

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

Table 29: Other External Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 7–12[a]

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
California Achievement Test (CAT)														
/ 4	Spr 98	Scaled	–	–	11	0.75*	–	–	–	–	–	–	–	–
/ 5	Spr 98	NCE	80	0.85*	–	–	–	–	–	–	–	–	–	–
Comprehensive Test of Basic Skills (CTBS)														
/ 4	Spr 97	NCE	–	–	12	0.68*	–	–	–	–	–	–	–	–
/ 4	Spr 98	NCE	43	0.84*	–	–	–	–	–	–	–	–	–	–
/ 4	Spr 98	Scaled	107	0.44*	15	0.57*	43	0.86*	–	–	–	–	–	–
A-16	Spr 98	Scaled	24	0.82*	–	–	–	–	–	–	–	–	–	–
Explore (ACT Program for Educational Planning, 8th Grade)														
	Fall 97	NCE	–	–	–	–	67	0.72*	–	–	–	–	–	–
	Fall 98	NCE	–	–	32	0.66*	–	–	–	–	–	–	–	–
Iowa Test of Basic Skills (ITBS)														
Form K	Spr 98	NCE	–	–	–	–	35	0.84*	–	–	–	–	–	–
Form K	Fall 98	NCE	32	0.87*	43	0.61*	–	–	–	–	–	–	–	–
Form K	Fall 98	Scaled	72	0.77*	67	0.65*	77	0.78*	–	–	–	–	–	–
Form L	Fall 98	NCE	19	0.78*	13	0.73*	–	–	–	–	–	–	–	–
Metropolitan Achievement Test (MAT)														
7th Ed.	Spr 97	Scaled	114	0.70*	–	–	–	–	–	–	–	–	–	–
7th Ed.	Spr 98	NCE	46	0.84*	63	0.86*	–	–	–	–	–	–	–	–
7th Ed.	Spr 98	Scaled	88	0.70*	–	–	–	–	–	–	–	–	–	–
7th Ed.	Fall 98	NCE	50	0.55*	48	0.75*	–	–	–	–	–	–	–	–
Missouri Mastery Achievement Test (MMAT)														
	Spr 98	Scaled	24	0.62*	12	0.72*	–	–	–	–	–	–	–	–
North Carolina End of Grade Test (NCEOG)														
	Spr 97	Scaled	–	–	–	–	–	–	58	0.81*	–	–	–	–
	Spr 98	Scaled	–	–	–	–	73	0.57*	–	–	–	–	–	–

Table 29: Other External Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 7–12[a] (Continued)

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
PLAN (ACT Program for Educational Planning, 10th Grade)														
	Fall 97	NCE	–	–	–	–	–	–	–	–	46	0.71*	–	–
	Fall 98	NCE	–	–	–	–	–	–	104	0.53*	–	–	–	–
Preliminary Scholastic Aptitude Test (PSAT)														
	Fall 98	Scaled	–	–	–	–	–	–	–	–	78	0.67*	–	–
Stanford Achievement Test (Stanford)														
9th Ed.	Spr 97	Scaled	–	–	–	–	–	–	–	–	–	–	11	0.90*
7th Ed.	Spr 98	Scaled	–	–	8	0.83*	–	–	–	–	–	–	–	–
8th Ed.	Spr 98	Scaled	6	0.89*	8	0.78*	91	0.62*	–	–	93	0.72*	–	–
9th Ed.	Spr 98	Scaled	72	0.73*	78	0.71*	233	0.76*	32	0.25	64	0.76*	–	–
4th Ed. 3/V	Spr 98	Scaled	–	–	–	–	–	–	55	0.68*	–	–	–	–
9th Ed.	Fall 98	NCE	92	0.67*	–	–	–	–	–	–	–	–	–	–
9th Ed.	Fall 98	Scaled	–	–	–	–	93	0.75*	–	–	–	–	70	0.75*
Stanford Reading Test														
3rd Ed.	Fall 97	NCE	–	–	–	–	5	0.81	24	0.82*	–	–	–	–
TerraNova														
	Fall 97	NCE	103	0.69*	–	–	–	–	–	–	–	–	–	–
	Spr 98	Scaled	–	–	87	0.82*	–	–	21	0.47*	–	–	–	–
	Fall 98	NCE	35	0.69*	32	0.74*	–	–	–	–	–	–	–	–
Test of Achievement and Proficiency (TAP)														
	Spr 97	NCE	–	–	–	–	–	–	–	–	36	0.59*	–	–
	Spr 98	NCE	–	–	–	–	–	–	41	0.66*	–	–	43	0.83*
Texas Assessment of Academic Skills (TAAS)														
	Spr 97	TLI	–	–	–	–	–	–	–	–	–	–	41	0.58*
Wide Range Achievement Test 3 (WRAT3)														
	Spr 98		9	0.35	–	–	–	–	–	–	–	–	–	–
	Fall 98		–	–	–	–	16	0.80*	–	–	–	–	–	–

Table 29: Other External Validity Data: STAR Reading 2 Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 7–12[a] (Continued)

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Wisconsin Reading Comprehension Test														
	Spr 98		–	–	–	–	–	–	63	0.58*	–	–	–	–
Summary														
Grade(s)		**All**	**7**		**8**		**9**		**10**		**11**		**12**	
Number of students		3,158	1,016		529		733		398		317		165	
Number of coefficients		60	18		15		10		8		5		4	
Average validity		–	0.71		0.72		0.75		0.60		0.69		0.77	
Overall average		0.71												

[a] Sample sizes are in the columns labeled “n.”

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

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

In the US, since the passage of the No Child Left Behind Act in 2001, all states have moved to comprehensive tests of grade level standards for purposes of accountability. This has created interest in the degree to which STAR Reading test scores are related to state accountability test scores. The following section provides specific information about the validity of STAR scores relative to state test scores. Results of concurrent and predictive validity (defined earlier) are presented here with specific results for a variety of state tests of accountability. This section will continually be updated as additional evidence of STAR score validity with respect to state tests is accumulated.

Tables 30 and 31 provide a variety of concurrent and predictive validity coefficients, respectively, for grades 3–8. Numerous state accountability tests have been used in this research.

Table 30: Concurrent Validity Data: STAR Reading 2 Correlations (r) with State Accountability Tests, Grades 3–8[a]

Date	Score	3		4		5		6		7		8	
		n	r	n	r	n	r	n	r	n	r	n	r
Colorado Student Assessment Program													
Spr 06	Scaled	82	0.75*	79	0.83*	93	0.68*	280	0.80*	299	0.84*	185	0.83*
Delaware Student Testing Program—Reading													
Spr 05	Scaled	104	0.57*	–	–	–	–	–	–	–	–	–	–
Spr 06	Scaled	126	0.43*	141	0.62*	157	0.59*	75	0.66*	150	0.72	–	–
Florida Comprehensive Assessment Test													
Spr 06	SSS	–	–	41	0.65*	–	–	–	–	–	–	74	0.65*
Illinois Standards Achievement Test—Reading													
Spr 05	Scaled	594	0.76*	–	–	449	0.70*	–	–	–	–	157	0.73*
Spr 06	Scaled	140	0.80*	144	0.80*	146	0.72*	–	–	140	0.70*	–	–
Michigan Educational Assessment Program—English Language Arts													
Fall 04	Scaled	–	–	155	0.81*	–	–	–	–	154	0.68*	–	–
Fall 05	Scaled	218	0.76*	196	0.80*	202	0.80*	207	0.69*	233	0.72*	239	0.70*
Fall 06	Scaled	116	0.79*	132	0.69*	154	0.81*	129	0.66*	125	0.79*	152	0.74*
Michigan Educational Assessment Program—Reading													
Fall 04	Scaled	–	–	155	0.80*	–	–	–	–	156	0.68*	–	–
Fall 05	Scaled	218	0.77*	196	0.78*	202	0.81*	207	0.68*	233	0.71*	239	0.69*
Fall 06	Scaled	116	0.75*	132	0.70*	154	0.82*	129	0.70*	125	0.86*	154	0.72*
Mississippi Curriculum Test													
Spr 02	Scaled	148	0.62*	175	0.66*	81	0.69*	–	–	–	–	–	–
Spr 03	Scaled	389	0.71*	359	0.70*	377	0.70*	364	0.72*	372	0.70*	–	–
Oklahoma Core Curriculum Test													
Spr 06	Scaled	78	0.62*	92	0.58*	46	0.52*	80	0.60*	–	–	–	–

Table 30: Concurrent Validity Data: STAR Reading 2 Correlations (r) with State Accountability Tests, Grades 3–8[a] (Continued)

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

[a] Sample sizes are in the columns labeled "n."

* Denotes correlation coefficients that are statistically significant ($p < 0.05$).

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

Predictor Date	Criterion Date[b]	3		4		5		6		7		8	
		n	r	n	r	n	r	n	r	n	r	n	r
Colorado Student Assessment Program													
Fall 05	Spr 06	82	0.72*	79	0.77*	93	0.70*	280	0.77*	299	0.83*	185	0.83*
Delaware Student Testing Program—Reading													
Fall 04	Spr 05	189	0.58*	–	–	–	–	–	–	–	–	–	–
Win 05	Spr 05	120	0.67*	–	–	–	–	–	–	–	–	–	–
Spr 05	Spr 06	161	0.52*	191	0.55*	190	0.62*	–	–	100	0.75*	143	0.63*
Fall 05	Spr 06	214	0.39*	256	0.62*	270	0.59*	242	0.71*	273	0.69*	247	0.70*
Win 05	Spr 06	233	0.47*	276	0.59*	281	0.62*	146	0.57*	–	–	61	0.64*
Florida Comprehensive Assessment Test													
Fall 05	Spr 06	–	–	42	0.73*	–	–	409	0.67*	381	0.61*	387	0.62*
Win 07	Spr 07	–	–	–	–	–	–	417	0.76*	342	0.64*	361	0.72*

Table 31: Predictive Validity Data: STAR Reading Scaled Scores Predicting Later Performance for Grades 3–8 on Numerous State Accountability Tests[a] (Continued)

Predictor Date	Criterion Date[b]	3		4		5		6		7		8	
		n	r	n	r	n	r	n	r	n	r	n	r
Illinois Standards Achievement Test—Reading													
Fall 04	Spr 05	450	0.73*	–	–	317	0.68*	–	–	–	–	–	–
Win 05	Spr 05	564	0.76*	–	–	403	0.68*	–	–	–	–	–	–
Fall 05	Spr 06	133	0.73*	140	0.74*	145	0.66*	–	–	173	0.51*	158	0.66*
Win 06	Spr 06	138	0.76*	145	0.77*	146	0.70*	–	–	–	–	–	–
Michigan Educational Assessment Program—English Language Arts													
Fall 04	Fall 05^{P}	193	0.60*	181	0.70*	170	0.75*	192	0.66*	181	0.71*	88	0.85*
Win 05	Fall 05^{P}	204	0.68*	184	0.74*	193	0.75*	200	0.70*	214	0.73*	212	0.73*
Spr 05	Fall 05^{P}	192	0.73*	171	0.73*	191	0.71*	193	0.62*	206	0.75*	223	0.69*
Fall 05	Fall 06^{P}	111	0.66*	132	0.71*	119	0.77*	108	0.60*	114	0.66*	126	0.66*
Win 06	Fall 06^{P}	114	0.77*	–	–	121	0.75*	109	0.66*	114	0.64*	136	0.71*
Spr 06	Fall 06^{P}	–	–	–	–	–	–	–	–	–	–	30	0.80*
Michigan Educational Assessment Program—Reading													
Fall 04	Fall 05^{P}	193	0.60*	181	0.69*	170	0.76*	192	0.66*	181	0.70*	88	0.84*
Win 05	Fall 05^{P}	204	0.69*	184	0.74*	193	0.78*	200	0.70*	214	0.72*	212	0.73*
Spr 05	Fall 05^{P}	192	0.72*	171	0.72*	191	0.74*	193	0.62*	206	0.72*	223	0.69*
Fall 05	Fall 06^{P}	111	0.63*	132	0.70*	119	0.78*	108	0.62*	116	0.68*	138	0.66*
Win 06	Fall 06^{P}	114	0.72*	–	–	121	0.75*	109	0.64*	116	0.68*	138	0.70*
Spr 06	Fall 06^{P}	–	–	–	–	–	–	–	–	–	–	30	0.81*
Mississippi Curriculum Test													
Fall 01	Spr 02	95	0.70*	97	0.65*	78	0.76*	–	–	–	–	–	–
Fall 02	Spr 03	337	0.67*	282	0.69*	407	0.71*	442	0.72*	425	0.68*	–	–
Oklahoma Core Curriculum Test													
Fall 04	Spr 05	–	–	–	–	44	0.63*	–	–	–	–	–	–
Win 05	Spr 05	–	–	–	–	45	0.66*	–	–	–	–	–	–
Fall 05	Spr 06	89	0.59*	90	0.60*	79	0.69*	84	0.63*	–	–	–	–
Win 06	Spr 06	60	0.65*	40	0.67*	–	–	–	–	–	–	–	–

Table 31: Predictive Validity Data: STAR Reading Scaled Scores Predicting Later Performance for Grades 3–8 on Numerous State Accountability Tests[a] (Continued)

Predictor Date	Criterion Date[b]	3		4		5		6		7		8	
		n	r	n	r	n	r	n	r	n	r	n	r
Summary													
Grades	**All**	**3**		**4**		**5**		**6**		**7**		**8**	
Number of students	22,018	4,493		2,974		4,086		3,624		3,655		3,186	
Number of coefficients	119	24		19		23		17		17		19	
Average validity	–	0.66		0.68		0.70		0.68		0.69		0.70	
Overall validity	0.68												

a. Grade given in the column signifies the grade within which the Predictor variable was given (as some validity estimates span contiguous grades).

b. ^{P} indicates a criterion measure was given in a subsequent grade from the predictor.

* Denotes significant correlation ($p < 0.05$).

Relationship of STAR Reading Enterprise Scores to Scores on Previous Versions ENTERPRISE

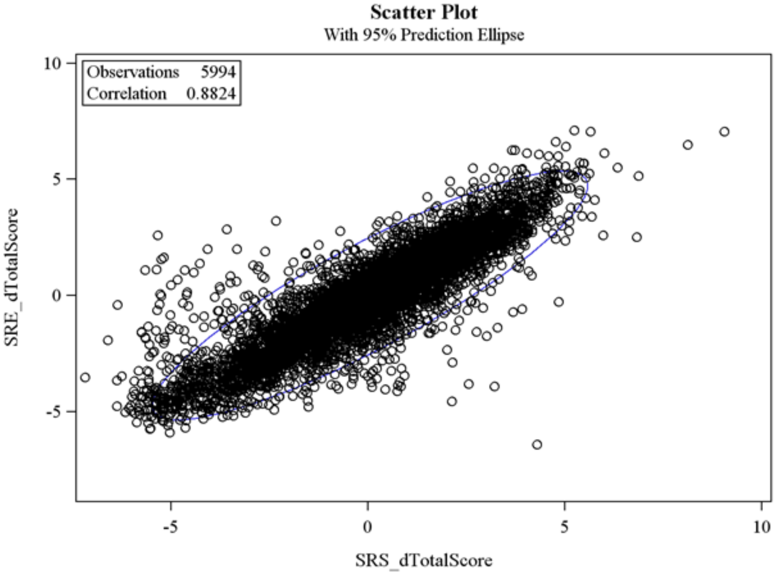
STAR Reading Enterprise represents a significant departure from previous versions of STAR. It is not a replacement for earlier versions; instead, it presents an alternative approach to reading assessment. Unlike previous STAR Reading versions, which were primarily designed as measures only of reading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

STAR Reading Enterprise was released for use in June 2011. To date, there has been little opportunity to collect data on the correlations between it and external reading test scores. However, in the course of its development, STAR Reading Enterprise was administered to thousands of students who also took previous versions. The correlations between STAR Reading Enterprise and previous versions of STAR Reading provide validity evidence of their own. To the extent that those correlations are high, they would provide evidence that Enterprise and previous versions are measuring the same or highly similar underlying attributes, even though they are dissimilar in content and measurement precision. Table 32 displays data on the correlations between STAR Reading Enterprise and scores on two previous versions: STAR Reading Classic (which includes versions 2.0 through 4.3) and STAR Reading Service (version 4.4 RP.) STAR Reading Classic and Service versions are highly similar to one another, differing primarily in terms of the software that delivers them; for all practical purposes, they may be considered alternate forms of STAR Reading.

Table 32: Correlations of STAR Reading Enterprise with Scores on STAR Reading Classic and STAR Reading Service Tests

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

Figure 3: Scatterplot of STAR Reading Enterprise and STAR Reading Service Test Scores for 5,994 Students Tested in June and July 2011



Meta-Analysis of the STAR Reading Validity Data

Meta-analysis is a statistical procedure for combining results from different sources or studies. When applied to a set of correlation coefficients that estimate test validity, meta-analysis combines the observed correlations and sample sizes to yield estimates of overall validity. In addition, standard errors and confidence intervals can be computed for overall validity estimates as well as within-grade validity estimates. To conduct a meta-analysis of the STAR Reading validity data, the 569 correlations reported in the current manual were combined and analyzed using a fixed-effects model for meta-analysis (see Hedges and Olkin, 1985, for a methodology description).

The results are displayed in Table 33. The table lists results for the correlations within each grade, as well as results from combining data from all twelve grades. For each set of results, the table lists an estimate of the true validity, a standard error, and the lower and upper limits of a 95 percent confidence interval for the validity coefficient. Using the 569 correlation coefficients, the overall estimate of the validity of STAR Reading is 0.78, with a standard error of 0.001. The 95 percent confidence interval allows one to conclude that the true validity coefficient for STAR Reading is approximately 0.78. The probability of observing the 569 correlations reported in Tables 24–27 if the true validity were zero, would be virtually zero. Because the 569 correlations were obtained with widely different tests, and among students from twelve different grades, these results provide strong support for the validity of STAR Reading as a measure of reading skills.

Table 33: Results of the Meta-Analysis of STAR Reading Correlations with Other Tests

	Effect Size		95% Confidence Level	
Grade	Validity Estimate	Standard Error	Lower Limit	Upper Limit
1	0.70	0.00	0.69	0.70
2	0.78	0.00	0.78	0.78
3	0.78	0.00	0.78	0.78
4	0.78	0.00	0.78	0.78
5	0.78	0.00	0.78	0.78
6	0.78	0.00	0.78	0.78
7	0.77	0.00	0.77	0.78
8	0.77	0.00	0.77	0.77
9	0.82	0.01	0.82	0.83
10	0.85	0.00	0.84	0.85
11	0.86	0.01	0.85	0.86
12	0.85	0.02	0.82	0.87
All	0.78	0.00	0.78	0.78

Post-Publication Study Data

Subsequent to publication of STAR Reading 2.0 in 1999, additional external validity data have become available, both from users of the assessment and from special studies conducted by Renaissance Learning and others. This section provides summaries of those new data along with tables of results. Data from four sources are presented here. They include a predictive validity study, a longitudinal study, a concurrent validity study in England, and a study of STAR Reading's construct validity as a measure of reading comprehension.

Predictive Validity: Correlations with SAT9 and the California Standards Tests

A doctoral dissertation (Bennicoff-Nan, 2002) studied the validity of STAR Reading as a predictor of student's scores in a California school district on the California Standards Test (CST) and the Stanford Achievement Tests, Ninth Edition (SAT9), the reading accountability tests mandated by the State of California. At the time of the study, those two tests were components of the California Standardized Testing and Reporting Program. The study involved analysis of test scores of more than

1,000 school children in four grades in a rural central California school district; 83 percent of students in the district were eligible for free and reduced lunch and 30 percent were identified as having limited English proficiency.

Bennicoff-Nan's dissertation addressed a number of different research questions. For purposes of this technical manual, we are primarily interested in the correlations between STAR Reading 2 with SAT9 and CST scores. Those correlations are displayed by grade in Table 34.

Table 34: Correlations of STAR Reading 2.0 Scores with SAT9 and California Standards Test Scores, by Grade

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

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

A Longitudinal Study: Correlations with SAT9

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

[2] Reading Renaissance is a supplemental reading program that uses STAR Reading and Accelerated Reader. STAR Reading scores help teachers match students with books at an appropriate difficulty level. Accelerated Reader encourages reading practice and monitors individual students' reading success on a daily basis.

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

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

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

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

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

[a] All correlations significant, p < 0.001.

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

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

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

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

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

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

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

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

[a] UK school year values are 1 greater than the corresponding US school grade. Thus, Year 2 corresponds to Grade 1, etc.

[b] Correlations with the individual SRS forms were calculated with within-form raw scores. The overall correlation was calculated with a vertical Scaled Score.

Construct Validity: Correlations with a Measure of Reading Comprehension

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

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

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

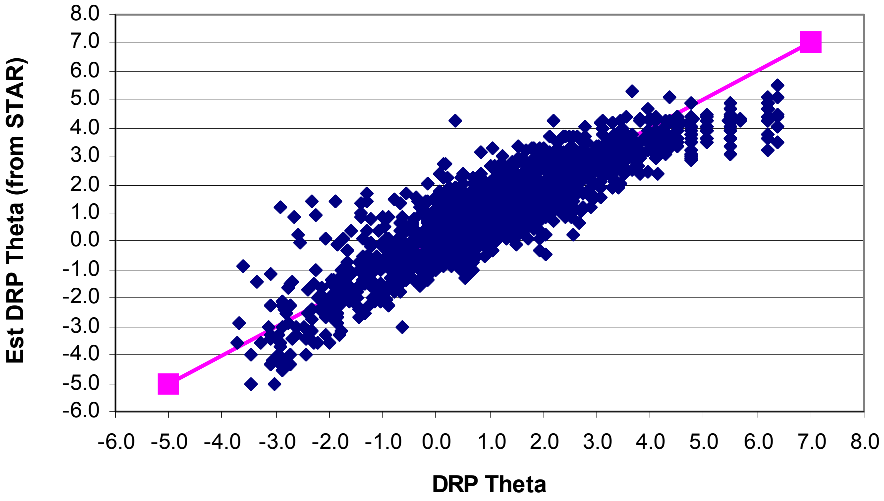
Subsequent to the factor analysis, the STAR Reading item difficulty parameters were transformed to the DRP difficulty scale, so that scores on both tests could be expressed on a common scale. STAR Reading scores on that scale were then calculated using the methods of Item Response Theory. The correlations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

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

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

Combining students across grade levels and plotting both their STAR Reading and DRP scores on the same yardstick yielded the plot as seen in Figure 4. The plot shows a slightly curvilinear relationship between STAR and DRP scales, but the strong linear correlation between scores on the two tests is evident as well.

Figure 4: STAR to DRP Linking Study Grades Combined (r = 0.89)



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

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

During the fall of 2007 and winter of 2008, 32 schools across the United States that were then using both STAR Reading and DIBELS oral reading fluency (DORF) for interim assessments were contacted and asked to participate in a research study. The schools were asked to ensure that students were tested during the fall and winter interim assessment schedules, usually during September and January, respectively, on both STAR Reading and DORF within a two-week time interval. Schools used the benchmark assessment passages from the grade-level-appropriate DORF passage sets.

In addition, schools were asked to submit data from the previous school year on the interim assessments. Any student that had a valid STAR Reading and DORF assessment within a two-week time span was used in the analysis. Thus, the research involved both a current sample of students who took benchmark assessments during the fall and winter of the 2007–2008 school year, as well as historical data from those same schools for students who took either the fall, winter, or spring benchmark assessments from the 2006–2007 school year.

This single-group design provided data for both evaluation of concurrent validity and the linking of the two score scales. For the linking analysis, an equipercentile methodology was used. Analysis was done independently for grades 1–4. Grade 1 data did not include any fall data, and all analyses were done using data from winter (both historical data from 2006–2007 and extant data collections during the 2007–2008 school year) and spring (historical data from the 2006–2007 school year). To evaluate the extent to which the linking accurately approximated student performance, 90 percent of the sample was used to calibrate the linking model, and the remaining 10 percent were used for cross-validating the results. The 10 percent were chosen by a simple random function.

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

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

The sample was composed of 61 percent of students of European ancestry; 21 percent of African ancestry; 11 percent of Hispanic ancestry; with the remaining 7 percent of Native American, Asian, or other ancestry. Just over 3 percent of the students were eligible for services due to limited English proficiency (LEP), and between 13 percent and 14 percent were eligible for special education services.

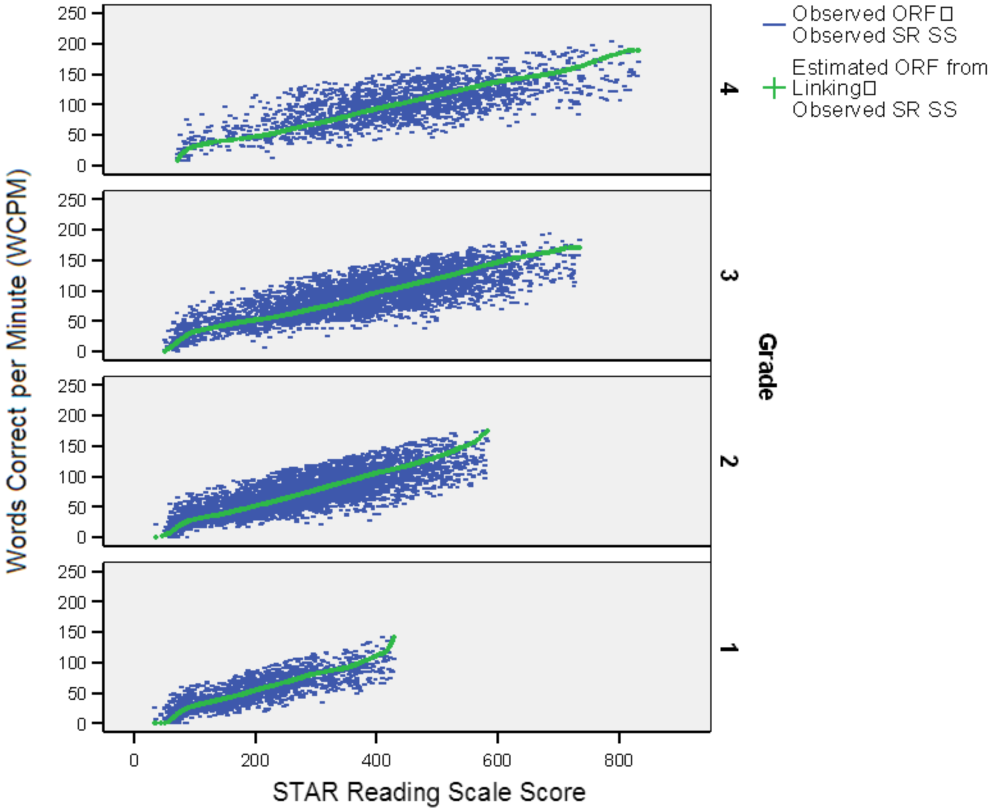
Students were individually assessed using the DORF benchmark passages. The students read the three benchmark passages under standardized conditions. The raw score for passages was computed as the number of words read correctly within the one-minute limit (WCPM, Words Correctly read Per Minute) for each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

Descriptive statistics for each grade in the study on STAR Reading Scaled Scores and DORF WCPM (words correctly read per minute) are found in Table 39. Correlations between the STAR Reading Scaled Score and DORF WCPM at all grades were significant ($p < 0.01$) and diminished consistently as grades increased. Figure 5 visualizes the scatterplot of observed DORF WCPM and SR Scaled Scores, with the equipercentile linking function overlaid. The equipercentile linking function appeared linear; however, deviations at the tails of the distribution for higher and lower performing students were observed. A table of selected STAR Reading Scaled Scores and corresponding Est. ORF values can be found in Appendix B on page 180. The root mean square error of linking for grades 1–4 was found to be 14, 19, 22, and 25, respectively.

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

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

Figure 5: Scatterplot of Observed DORF WCPM and SR Scale Scores for Each Grade with the Grade Specific Linking Function Overlaid



Cross-Validation Study Results

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

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

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

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

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

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

Summary of STAR Reading Validity Data

The validity data presented in this manual includes evidence of STAR Reading's concurrent, retrospective, predictive, and construct validity. The Meta-Analysis section showed the average uncorrected correlation between STAR Reading and all other reading tests to be 0.78. (Many meta-analyses adjust the correlations for range restriction and attenuation to less than perfect reliability; had we done that here, the average correlation would have exceeded 0.85.) Correlations with specific measures of reading ability were often higher than this average. For example, Bennicoff-Nan (2002) found very consistent within-grade correlations averaging 0.82 with SAT9 and 0.80 with the California Standards Test. Yoes (1999) found within-grade correlations with DRP averaging 0.81. When these data were combined across grades, the correlation was 0.89. The latter correlation may be interpreted as an estimate of the construct validity of STAR Reading as a measure of reading comprehension. Yoes also reported that results of item factor analysis of DRP and STAR Reading items yielded a single common factor. This provides strong support for the assertion that STAR Reading is a measure of reading comprehension.

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

Finally, the data showing the relationship between the new, longer, standards-based STAR Reading Enterprise test and scores on the traditional, shorter STAR Reading measures of reading comprehension alone, show that the correlation of the traditional tests with the new Enterprise version is nearly as high as the correlation between traditional STAR Reading and external measures of reading comprehension. While more data need to be brought to bear on this, it appears that despite their appreciable differences in content and measurement precision, the traditional and standards-based versions of STAR Reading are arguably measuring the same underlying reading ability attribute.

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

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

For many teachers, this is an informal process in which classroom assessments are used to monitor student performance on state standards. While this may be helpful, such assessments may be technically inadequate when compared to more standardized measures of student performance. Recently the assessment scale associated with STAR Reading has been linked to the scales used for summative reading tests in approximately 30 states, a number that is expected to increase in the near future. Linking STAR Reading assessments to state tests allows educators to reliably predict student performance on their state assessment using STAR Reading scores. More specifically, it places teachers in a position to identify

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

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

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

Methodology Comparison

Recently, Renaissance Learning has developed linkages between STAR Reading Scaled Scores and scores on the accountability tests of a number of states. Depending on the kind of data available for such linking, these linkages have been accomplished using one of two different methods. One method used student-level data, where both STAR and state test scores were available for the same students. The other method used school-level data; this method was applied when approximately 100% of students in a school had taken STAR Reading, but individual students' state test scores were not available.

Student-Level Data

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

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

School-Level Data

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

School-level data were analyzed for some of the states included in the student-level linking analysis. In an effort to increase sample size, the school-level data presented here represent "projected" Scaled Scores. Each STAR score was projected to the mid-point of the state test administrations window using decile-based growth norms. The growth norms are both grade- and subject-specific and are based on the growth patterns of more than one million students using STAR assessments over a three-year period. Again, the linking process used for school-level data is very similar to the previously described process—the distribution of state test scores is compared to projected STAR scores and using the observed distribution of state-test scores, equivalent cut scores are created for the STAR assessments (the key difference being that these comparisons are made at the group level).

Accuracy Comparisons

Accuracy comparisons between student- and school-level data are particularly important given the marked resource differences between the two methods. These comparisons are presented for three states[3] in Tables 42–44. With few exceptions, results of linking using school-level data were nearly identical to student-level data on measures of specificity, sensitivity, and overall accuracy. McLaughlin and Bandeira de Mello (2002) employed similar methods in their comparison of NAEP scores and state assessment results, and this method has been used several times since then (McLaughlin & Bandeira de Mello, 2003; Bandeira de Mello, Blankenship, & McLaughlin, 2009; Bandeira et al., 2008).

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

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

Table 42: Number of Students Included in Student-Level and School-Level Linking Analyses by State, Grade, and Subject

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

Table 43: Comparison of School Level and Student Level Classification Diagnostics for Reading/Language Arts

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

a. Sensitivity refers to the proportion of correct positive predictions.

b. Specificity refers to the proportion of negatives that are correctly identified (e.g. student will not meet a particular cut score).

c. False + rate refers to the proportion of students incorrectly identified as "at-risk."

d. False – rate refers to the proportion of students incorrectly identified as *not* "at-risk."

Table 44: Comparison of Differences Between Achieved and Forecasted Performance Levels in Reading/Language Arts (Forecast % – Achieved %)

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

The National Center on Response to Intervention (NCRTI) and Screening

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

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

- classification accuracy
- validity
- disaggregated validity and classification data for diverse populations

NCRTI ratings include four qualitative labels: convincing evidence, partially convincing evidence, unconvincing evidence, or data unavailable/inadequate. Please refer to Table 45 for descriptions of these labels as provided by NCRTI, as well as the scores assigned to STAR Reading in each of the categories. Further descriptive information is provided within the following tables.

Table 45: NCRTI Screening Indicator Descriptions

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

Aggregated Classification Accuracy Data

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

> “Receiver Operating Characteristic (ROC) curves are a useful way to interpret sensitivity and specificity levels and to determine related cut scores. ROC curves are a generalization of the set of potential combinations of sensitivity and specificity possible for predictors.” (Pepe, Janes, Longton, Leisenring, & Newcomb, 2004)
>
> “ROC curve analyses not only provide information about cut scores, but also provide a natural common scale for comparing different predictors that are measured in different units, whereas the odds ratio in logistic regression analysis must be interpreted according to a unit increase in the value of the predictor, which can make comparison between predictors difficult.” (Pepe, et al., 2004)
>
> “An overall indication of the diagnostic accuracy of a ROC curve is the area under the curve (AUC). AUC values closer to 1 indicate the screening measure reliably distinguishes among students with satisfactory and unsatisfactory

> reading performance, whereas values at .50 indicate the predictor is no better than chance." (Zhou, Obuchowski & Obushcowski, 2002)

Brief Description of the Current Sample and Procedure

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

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

An ROC analysis was used to compare the performance data on STAR Reading to performance data on state achievement tests. The STAR Reading Scaled Scores used for analysis originated from assessments 3–11 months before the state achievement test was administered. Selection of cut scores was based on the graph of sensitivity and specificity versus the Scaled Score. For each grade, the Scaled Score chosen as the cut point was equal to the score where sensitivity and specificity intersected. The classification analyses, cut points and outcome measures are outlined in Table 46. When collapsed across ethnicity, AUC values were all greater than 0.80. Descriptive notes for other values represented in the table are provided in the table footnote.

Table 46: Classification Accuracy in Predicting Proficiency on State Achievement Tests in Seven States[a]

Statistic[b]	Initial Analysis	Secondary Analysis
	Value	Value
False Positive Rate	0.2121	0.1824
False Negative Rate	0.2385	0.2201
Sensitivity	0.7615	0.7799

Table 46: Classification Accuracy in Predicting Proficiency on State Achievement Tests in Seven States[a] (Continued)

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

[a] Arkansas, Delaware, Illinois, Kansas, Michigan, Mississippi, and South Dakota.

[b] The false positive rate is equal to the proportion of students incorrectly labeled “at-risk.” The false negative rate is equal to the proportion of students incorrectly labeled not “at-risk.” Likewise, sensitivity refers to the proportion of correct positive predictions while specificity refers to the proportion of negatives that are correctly identified (e.g., student will not meet a particular cut score).

Aggregated Validity Data

Table 47 provides aggregated validity values as well as concurrent and predictive validity evidence for STAR Reading. Median validity coefficients ranged from 0.68–0.84.

Table 47: Overall Concurrent and Predictive Validity Evidence for STAR Reading

Type of Validity	Grade	Test	N (Range)	Coefficient	
				Range	Median
Predictive	3–6	CST	1,000+	0.78–0.81	0.80
Predictive	2–6	SAT9	44–389	0.66–0.73	0.68
Concurrent	1–8	Suffolk Reading Scale	2,694	0.78–0.88	0.84
Construct	3, 5, 7, 10	DRP	273–424	0.76–0.86	0.82
Concurrent	1–4	DIBELS Oral Reading Fluency	12,220	0.71–0.87	0.81
Predictive	1–6	State Achievement Tests	74,877–200,929	0.68–0.82	0.79
Predictive	7–12	State Achievement Tests	3,107–64,978	0.81–0.86	0.82
Concurrent	3–8	State Achievement Tests	1,200–2,329	0.71–0.74	0.73
Predictive	3–8	State Achievement Tests	2,974–4,493	0.66–0.70	0.68

Disaggregated Validity and Classification Data

Table 48 shows the disaggregated classification accuracy data for ethnic subgroups and also the disaggregated validity data.

Table 48: Disaggregated Classification and Validity Data

Classification Accuracy in Predicting Proficiency on State Achievement Tests in 6 States (Arkansas, Delaware, Illinois, Kansas, Michigan, and Mississippi): by Race/Ethnicity					
	White, non-Hispanic (n = 17,567)	Black, non-Hispanic (n = 8,962)	Hispanic (n = 1,382)	Asian/Pacific Islander (n = 231)	American Indian/Alaska Native (n = 111)
False Positive Rate	0.3124	0.4427	0.3582	0.1710	0.1216
False Negative Rate	0.3762	0.1215	0.1224	0.2368	0.4054
Sensitivity	0.6238	0.8785	0.8776	0.7632	0.5946
Specificity	0.8676	0.5573	0.6418	0.8290	0.8784
Positive Predictive Power	0.5711	0.5031	0.6103	0.4677	0.7097
Negative Predictive Power	0.8909	0.8999	0.8913	0.9467	0.8125
Overall Classification Rate	0.8139	0.6658	0.7337	0.8182	0.7838

Table 48: Disaggregated Classification and Validity Data (Continued)

Classification Accuracy in Predicting Proficiency on State Achievement Tests in 6 States (Arkansas, Delaware, Illinois, Kansas, Michigan, and Mississippi): by Race/Ethnicity										
	White, non-Hispanic (n = 17,567)		**Black, non-Hispanic (n = 8,962)**		**Hispanic (n = 1,382)**		**Asian/Pacific Islander (n = 231)**		**American Indian/Alaska Native (n = 111)**	
AUC (ROC)	**Grade**	**AUC**	**Grade**	**AUC**	**Grade**	**AUC**	**Grade**	**AUC**	**Grade**	**AUC**
	2	n/a	2	0.500	2	n/a	2	n/a	2	n/a
	3	0.863	3	0.828	3	0.868	3	0.913	3	0.697
	4	0.862	4	0.823	4	0.837	4	0.869	4	0.888
	5	0.853	5	0.832	5	0.839	5	0.855	5	0.919
	6	0.849	6	0.806	6	0.825	6	0.859	6	0.846
	7	0.816	7	0.784	7	0.866	7	0.904	7	0.900
	8	0.850	8	0.827	8	0.812	8	0.961	8	1.000
	9	1.000	9	0.848	9	n/a	9	n/a	9	n/a
	10	0.875	10	0.831	10	0.833	10	n/a	10	n/a
	11	0.750	11	1.000	11	n/a	11	n/a	11	n/a
Base Rate	0.2203		0.3379		0.3900		0.1645		0.333	
Cut Scores	**Grade**	**Cut Score**	**Grade**	**Cut Score**	**Grade**	**Cut Score**	**Grade**	**Cut Score**	**Grade**	**Cut Score**
	2	228	2	228	2	228	2	228	2	228
	3	308	3	308	3	308	3	308	3	308
	4	399	4	399	4	399	4	399	4	399
	5	488	5	488	5	488	5	488	5	488
	6	540	6	540	6	540	6	540	6	540
	7	598	7	598	7	598	7	598	7	598
	8	628	8	628	8	628	8	628	8	628
	9	708	9	708	9	708	9	708	9	708
	10	777	10	777	10	777	10	777	10	777
	11	1,055	11	1,055	11	1,055	11	1,055	11	1,055

Disaggregated Validity					
				Coefficient	
Type of Validity	**Age or Grade**	**Test or Criterion**	**n (range)**	**Range**	**Median**
Predictive (White)	2–6	SAT9	35–287	0.69–0.75	0.72
Predictive (Hispanic)	2–6	SAT9	7–76	0.55–0.74	0.675

The National Center on Intensive Intervention (NCII) and Progress Monitoring

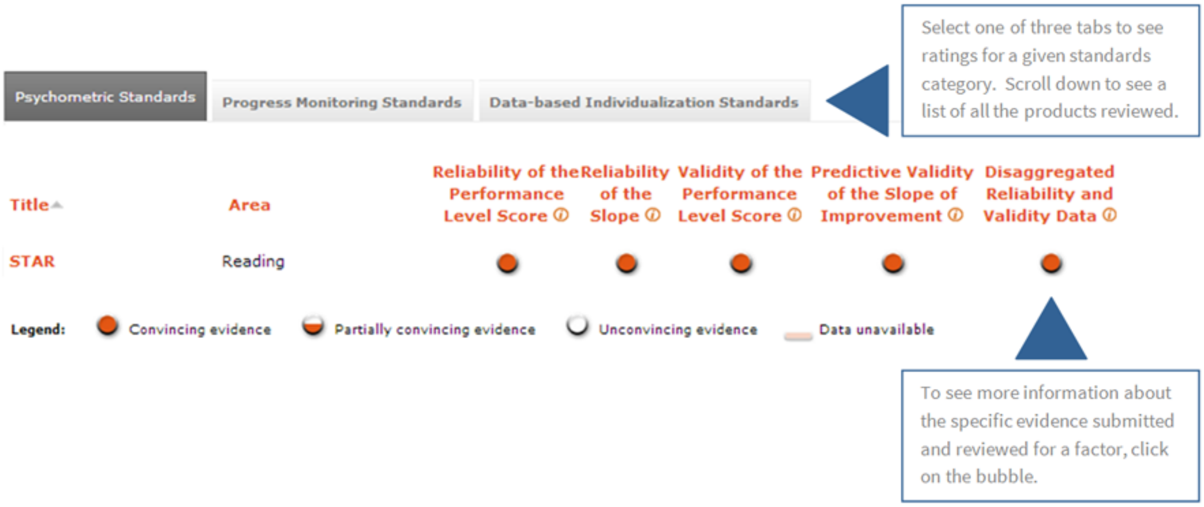
NCII is a more recent federally funded project; it is related to NCRTI but was created in 2012 with a mission focusing on just those students with severe learning needs. NCII reviews are currently ongoing and focus on the technical adequacy of assessments as progress-monitoring tools. The technical criteria and rating system were carried over from NCII, and STAR Reading has again earned strong ratings compared to other reading assessments.

When evaluating progress monitoring tools, NCII considers a variety of factors in three general standards categories:

- Psychometric Standards
- Progress Monitoring Standards
- Data-Based Individualization Standards

Please refer to the NCII website for the most up to date information about the factors included in reviews and scores assigned to STAR Reading: http://www.intensiveintervention.org/chart/progress-monitoring. Figure 6 provides a snapshot of the NCII website navigation features.

Figure 6: Screenshot from NCII Website: Academic Progress Monitoring General Outcome Measures


Psychometric Standards
Progress Monitoring Standards
Data-based Individualization Standards

Title▲	Area	Reliability of the Performance Level Score ⓘ	Reliability of the Slope ⓘ	Validity of the Performance Level Score ⓘ	Predictive Validity of the Slope of Improvement ⓘ	Disaggregated Reliability and Validity Data ⓘ
STAR	Reading	●	●	●	●	●

Legend: ● Convincing evidence ◐ Partially convincing evidence ○ Unconvincing evidence ▭ Data unavailable

Select one of three tabs to see ratings for a given standards category. Scroll down to see a list of all the products reviewed.

To see more information about the specific evidence submitted and reviewed for a factor, click on the bubble.

Norming

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

Test Score Norms

National norms for STAR Reading version 1 were collected in 1996. Substantial changes introduced in STAR Reading version 2 necessitated the development of new norms in 1999. Those norms were used until new norms were developed in 2008. The 2008 norms were used until the 2014–2015 school year, when they were updated. This section describes the development of those updated norms, which was completed in mid-2014.

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

This chapter describes the development of the 2014 norms, using data collected over the course of two full school years: 2011–2012 and 2012–2013. Prior to 2008, norms were developed by means of special-purpose norming studies, in which national samples of schools were cast, and those schools were solicited to participate in the norming by administering a special norming version of the assessment. The spring 2014 norming of STAR Reading is the second instance in which national samples of students were drawn from routine administrations of STAR Reading. Details of the procedures employed are given in this chapter.

Students participating in the norming study took assessments between September 1, 2011 and June 30, 2013. Students took the STAR Reading tests under normal test administration conditions. No specific norming test was developed and no deviations were made from the usual test administration. Thus, students in the norming sample took STAR Reading tests as they are administered in everyday use.

Sample Characteristics

During the norming period, a total of 5,240,114 US students in grades 1–12 took STAR Reading Enterprise tests administered using Renaissance Place servers hosted by Renaissance Learning.

The first step in sampling was to select a matched sample of only students who had tested in both the fall and the spring of either the 2011–2012 or 2012–2013 school years. Some students tested in both school years while others tested in only one school year. From the matched sample, a stratified subsample was randomly drawn based on student grade and ability decile. Ability decile was determined by grouping students in each grade into deciles based on their spring Rasch ability scores in STAR Reading Enterprise. The grade and decile sampling was necessary to ensure adequate and similar distribution of students in each decile and grade. Because these norming data were a convenience sample drawn from the STAR Reading customer base, steps were taken to ensure the resulting norms were nationally representative of grades 1–12 US student population with regard to certain important characteristics. A post-stratification procedure was used to adjust the sample proportions to the approximate national proportions on three key variables: geographic region, district socio-economic status, and district/school size. These three variables were chosen because they had previously been used in STAR Reading norming studies to draw nationally representative samples, are known to be related to test scores, and were readily available for the schools in the Renaissance Place hosted database.

The final norming sample size, after selecting only students with test scores in both the fall and the spring of each year in the norming period and further sampling by grade and ability decile was 1,321,800 students in grades 1–12; some students were counted in two grades if they had test records in both of the school years sampled in the 2014 norms. The unique student count was 1,188,610. These students came from 9,768 schools across 50 states and the District of Columbia.

Table 49 provides a breakdown of the number of students participating per grade.

Table 49: Numbers of Students per Grade in the Norms Sample

Grade	N	Grade	N	Grade	N
1	54,570	5	214,390	9	15,720
2	288,910	6	96,130	10	16,380
3	270,570	7	61,100	11	12,320
4	250,200	8	35,040	12	6,470
				Total	1,321,800

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

- National population estimates of students' demographics (Ethnicity and gender) in grades 1–12 were obtained from NCES; these estimates were from 2011, the most recent data available. National estimates of race/ethnicity were computed using the NCES data based on single race/ethnicity and also a multiple race category. The NCES data reflect the most recent census data from the US census bureau.
- National estimates of school-related characteristics were obtained from November 2013 Market Data Retrieval (MDR) information. The MDR database contains the most recent data on schools, some of which may not be reflected in the NCES data.

Table 50 shows national percentages of children in grades 1–12 by region, school/district enrollment, district socio-economic status, and location, along with the corresponding percentages in the sample summarized in Table 49. MDR estimates of geographic region were based on the four broad areas identified by the National Educational Association as Northeastern, Midwestern, Southeastern, and Western regions. The specific states in each region are shown below.

Northeast:

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

Midwest:

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

Southeast:

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

West:

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

District socioeconomic status (SES) was defined by the percent of students within the district that were eligible for free/reduced price lunches (as defined by NCES) and was based on students attending both public and non-public schools. School enrollment was defined as the average number of students per school within the district. School location was defined as urban, suburban, town, or rural using the definitions utilized by MDR.

Table 50: Sample Characteristics Along with National Population Estimates and Sample Estimates

		National Estimates	Norming Sample
Region	Midwest	21.5%	21.9%
	Northeast	19.2%	12.0%
	Southeast	24.4%	33.2%
	West	34.9%	33.0%
School Enrollment	< 200	4.6%	3.1%
	200–499	27.6%	39.0%
	≥ 500	67.9%	57.9%
District Socioeconomic Status	Low	17.7%	24.2%
	Below Median	23.3%	29.4%
	Above Median	25.5%	25.3%
	High	33.5%	21.0%
Location	Rural	19.3%	27.2%
	Suburban	36.3%	29.1%
	Town	12.1%	15.9%
	Urban	32.3%	27.9%

Table 55 provides information on the demographic characteristics of students in the sample and national percentages provided by NCES. No weighting was done on the basis of these demographic variables; they are provided to help describe the sample of students and the schools they attended. School type was defined to be either public (including charter schools) or non-public (private, Catholic).

Table 51: Student Demographics and School Information: National Estimates and Sample Percentages

			National Estimate	Norming Sample
Gender	Public	Female	48.6%	50.2%
		Male	51.4%	49.8%
	Non-Public	Female	–	51.2%
		Male	–	48.8%
Race/Ethnicity	Public	American Indian	1.1%	1.6%
		Asian	5.0%	3.7%
		Black	16.0%	22.9%
		Hispanic	23.1%	19.8%
		White	52.5%	52.0%
		Multiple Race[a]	2.4%	–
	Non-Public	American Indian	0.4%	8.1%
		Asian	5.7%	4.0%
		Black	9.3%	9.4%
		Hispanic	9.7%	29.7%
		White	72.2%	48.8%
		Multiple Race[a]	2.7%	–

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

Test Administration

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

Data Analysis

Student test records were compiled from the complete database of STAR Reading Renaissance Place users. Data spanned two school years from September 2011 to June 2013. Students' Rasch scores on their first STAR Reading Enterprise test taken between the first and the third month of the school year based on grade placement were used to compute norms for the fall; students' Rasch scores on the last STAR Reading Enterprise test taken between the 7th and the 9th month of the school year were used to compute norms for the spring. Interpolation was used to estimate norms for times of the year between the first month in the fall and the last month in the spring. The norms were based on the distribution of Rasch scores for each grade.

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

Because norm-referenced scores such as percentile ranks and grade equivalents (GE) are specific to the normative sample, those scores should not be compared between the previous versions of STAR Reading and versions that employ the 2014 norms. If it is necessary to track student change across time and the new norms interrupt that tracking, it is necessary to use the scaled score, as that metric has not changed and the unit has remained the same. In addition, it is inadvisable to continue to use the older norms, which were collected in 2008, as the newer norms collected in 2014 represent more current estimates of the population of US school children.

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

		Fall Scaled Scores			Spring Scaled Scores		
Grade	N	Mean	Standard Deviation	Median	Mean	Standard Deviation	Median
1	54,570	105	68	77	204	112	185
2	288,910	238	126	224	355	151	341
3	270,570	366	152	357	464	178	447
4	250,200	470	177	458	563	214	528
5	214,390	589	220	560	688	260	639
6	96,130	716	270	671	816	303	792
7	61,100	811	304	795	898	320	893
8	35,040	914	318	908	987	320	991
9	15,720	986	321	989	1041	313	1101
10	16,380	1022	321	1085	1064	312	1161
11	12,320	1058	308	1152	1087	300	1189
12	6,470	1103	304	1219	1116	300	1235

Growth Norms

To enhance the utility of STAR assessments for indexing growth, two types of growth metrics are calculated annually: Student Growth Percentile (Time-Adjusted Model) (SGP (TAM)) and growth norms. Both are norm-referenced estimates that compare a student's growth to that of his or her academic peers nationwide. SGP (TAM)s use 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. SGP (TAM)s 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 SGP (TAM)s 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 Reading are based on over 3 million student assessments (N = 3,518,976). 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 Reading. 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 Reading 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 Definitions

This chapter enumerates all of the scores reported by STAR Reading, including scaled scores, norm-referenced, and criterion-referenced scores. In order to calculate the norm-referenced scores correctly, STAR Reading must have accurate information about each student's grade placement. Below is an extensive discussion of the importance of having the correct grade placement for each student. Definitions and descriptions of the score scales follow the grade placement discussion.

Types of Test Scores

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

- *Criterion-referenced scores* describe a student's performance relative to a specific content domain or to a standard. Such scores may be expressed either on a continuous score scale or as a classification. An example of a criterion-referenced score on a continuous scale is a percent-correct score, which expresses what proportion of test questions the student can answer correctly in the content domain. An example of a criterion-referenced classification is a proficiency category on a standards-based assessment: the student may be said to be "proficient" or not, depending on whether the student's score equals, exceeds, or falls below a specific criterion (the "standard") used to define "proficiency" on the standards-based test. The criterion-referenced score reported by STAR Reading is the Instructional Reading Level, which compares a student's test performance to the 1995 updated vocabulary lists that are based on the EDL's Core Vocabulary list. The Instructional Reading Level is the highest grade level at which the student is estimated to comprehend 80 percent of the text written at that level.
- *Norm-referenced scores* compare a student's test results to the results of other students who have taken the same test. In this case, scores provide a relative measure of student achievement compared to the performance of a group of students at a given time. Percentile Ranks and Grade Equivalents are the two primary norm-referenced scores available in STAR Reading software. Both of these scores are based on a comparison of a student's test results to the data collected during the 2014 national norming program.

Scaled Score (SS)

STAR Reading software creates a virtually unlimited number of test forms as it dynamically interacts with the students taking the test. In order to make the results of all tests comparable, and in order to provide a basis for deriving the norm-referenced scores, it is necessary to convert all the results of STAR Reading tests to scores on a common scale. STAR Reading 2 and higher software does this in two steps. First, maximum likelihood is used to estimate each student's location on the Rasch ability scale, based on the difficulty of the items administered and the pattern of right and wrong answers. Second, the Rasch ability scores are converted to STAR Reading Scaled Scores, using the conversion table described in "Item and Scale Calibration" on page 32. STAR Reading 2 and higher Scaled Scores range from 0–1400.

Grade Equivalent (GE)

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

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

STAR Reading Grade Equivalents range from 0.0–12.9+. The scale divides the academic year into 10 monthly increments, and is expressed as a decimal with the unit denoting the grade level and the individual "months" in tenths. Table 53 indicates how the GE scale corresponds to the various calendar months. For example, if a student obtained a GE of 4.6 on a STAR Reading assessment, this would suggest that the student was performing similarly to the average student in the fourth grade at the sixth month (March) of the academic year. Because the STAR Reading 4.x norming took place during the end of the seventh month (September) and the entire eighth month of the school year (May), the GEs ending in .8 are empirically based, and based on the observed data from the normative sample. All other monthly GE scores are derived through interpolation by fitting a curve to the grade-by-grade medians. Table 57 on page 153 contains the Scaled Score to GE conversions.

Table 53: Incremental Grade Placements per Month

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

[a] Depends on the current school year set in Renaissance Place.

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

Estimated Oral Reading Fluency (Est. ORF)

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

Est. ORF is reported as an estimated number of words a student can read correctly within a one-minute time span on grade-level-appropriate text. Grade-level text was defined to be connected text in a comprehensible passage form that has a readability level within the range of the first half of the school year. For instance, an Est. ORF score of 60 for a second-grade student would be interpreted as meaning the student is expected to read 60 words correctly within one minute on a passage with a readability level between 2.0 and 2.5. Therefore, when this estimate is compared to an observed score on a specific passage, which has a fixed level of readability, there might be noticeable differences as the Est. ORF provides an estimate across a range of readability levels.

The Est. ORF score was computed using the results of a large-scale research study investigating the linkage between the STAR Reading scores and estimates of oral reading fluency on a range of passages with grade-level-appropriate difficulty. An equipercentile linking was done between STAR Reading scores and oral reading fluency providing an estimate of the oral reading fluency for each Scaled Score unit in STAR Reading for grades 1–4 independently. Results of the analysis can be found in “Post-Publication Study Data” on page 90. A table of selected STAR Reading Scaled Scores and corresponding Est. ORF values can be found in Appendix B on page 180.

Comparing the STAR Reading Test with Classical Tests

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

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

Instructional Reading Level (IRL)

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

As a construct, instructional reading levels have existed in the field of reading education for over fifty years. During this time, a variety of assessment instruments have been developed using different measurement criteria that teachers can use to estimate IRL. STAR Reading software determines IRL scores relative to 1995 updated vocabulary lists that are based on the Educational Development Laboratory's (EDL) *A Revised Core Vocabulary* (1969). The Instructional Reading Level is defined as the highest reading level at which the student can read at 90–98 percent word recognition (Gickling & Haverape, 1981; Johnson, Kress & Pikulski, 1987; McCormick, 1999) and with 80 percent comprehension or higher (Gickling & Thompson, 2001). Although STAR Reading does not directly assess word recognition, STAR Reading 2 and higher uses the student's Rasch ability scores, in conjunction with the Rasch difficulty parameters of graded vocabulary items, to determine the proportion of items a student can comprehend at each grade level.

Special IRL Scores

If a student's performance on STAR Reading 2 or 3 RP and higher indicates an IRL below the first grade, STAR Reading software will automatically assign an IRL score of Primer (P) or Pre-Primer (PP). Because the kindergarten-level test items are designed so that even readers of very early levels can understand them, a Primer or Pre-Primer IRL means that the student is essentially a non-reader. There are, however, other unusual circumstances that could cause a student to receive an IRL of Primer or Pre-Primer. Most often, this happens when a student simply does not try or purposely answers questions incorrectly.

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

Understanding IRL and GE Scores

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

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

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

In effect, the IRL references each student's STAR Reading performance to the difficulty of written material appropriate for instruction. This is a valuable piece of information in planning the instructional program for individuals or groups of students.

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

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

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

Percentile Rank (PR)

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

The Percentile Rank scale is not an equal-interval scale. For example, for a student with a grade placement of 7.7, a Scaled Score of 1,119 corresponds to a PR of 80, and a Scaled Score of 1,222 corresponds to a PR of 90. Thus, a difference of 103 Scaled Score points represents a 10-point difference in PR. However, for the same student, a Scaled Score of 843 corresponds to a PR of 50, and a Scaled Score of 917 corresponds to a PR of 60. While there is now only a 74-point difference in Scaled Scores, there is still a 10-point difference in PR. For this reason, PR scores should not be averaged or otherwise algebraically manipulated. NCE scores are much more appropriate for these activities.

Table 58 on page 158 contains an abridged version of the Scaled Score to Percentile Rank conversion table that the STAR Reading software uses. The actual table includes data for all of the monthly grade placement values from 1.0–12.9. Because STAR Reading norming occurred in the seventh month of the school year (May), the values for each grade are empirically based. The remaining monthly values were estimated by interpolating between the empirical points. Table 58 contains the interpolated norms for month 0 (zero) of the school year.

This table can be used to estimate PR values for tests that were taken when the grade placement value of a student was incorrect (see "Types of Test Scores" on page 120 for more information). If the error is caught right away, one always has the option of correcting the grade placement for the student and then having the student retest. However, the correction technique using this table, illustrated below in example form, is intended to provide an alternate correction procedure that does not require retesting.

If a grade placement error occurred because a third-grade student who tested in February was for some reason registered as a fourth-grader, his or her Percentile Rank and NCE scores will be in considerable error. In order to obtain better estimates of this student's norm-referenced scores, the educator will need to enter Table 58 in the 3.0 grade placement column and proceed down the table until the student's Scaled Score (or the next-higher value) is found in the table. Then, the educator will need to read off the left side of the table the PR value associated with this particular Scaled Score for a student at the beginning of the third grade. Next, the educator will need to follow the same procedure using the 4.0 grade placement column to obtain a PR corresponding to the same Scaled Score, had the student been at the beginning of the fourth grade. Then the educator will need to average the two PR values to obtain a better estimate of the student's PR (averaged because February is in the middle of the school year).

Teachers can use a similar interpolation procedure to obtain PR values that correspond to scores that would have been obtained at other times throughout the school year. This procedure, however, is only an approximation technique designed to compensate for grossly incorrect scores that result from a student testing while his or her grade placement was incorrectly specified. A slightly better technique involves finding the PR values in Table 58 (page 158), converting them to NCE values using Table 59 (page 162), interpolating between the NCE values, and then converting the interpolated NCE value back to a PR value using Table 60 (page 163).

Normal Curve Equivalent (NCE)

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

Table 59 on page 162 provides the NCEs corresponding to integer PR values and facilitates the conversion of PRs to NCEs. Table 60 on page 163 provides the conversions from NCE to PR. The NCE values are given as a range of scores that convert to the corresponding PR value.

Student Growth Percentile (Time-Adjusted Model) (SGP (TAM))

Student Growth Percentile (Time-Adjusted Model) (SGP (TAM))s are a norm-referenced quantification of individual student growth derived using quantile regression techniques. An SGP (TAM) compares a student’s growth to that of his or her academic peers nationwide. SGP (TAM)s provide a measure of how a student changed from one STAR testing window[4] to the next relative to other students with similar starting STAR Reading scores. SGP (TAM)s 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 (TAM) of 70 means that the student’s growth from one test to

[4] 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.

another exceeds the growth of 70% of students nationwide in the same grade with a similar beginning (pretest) STAR Reading score. All students, no matter their starting STAR score, have an equal chance to demonstrate growth at any of the 99 percentiles.

SGP (TAM)s are often used to indicate whether a student's growth is more or less than can be expected. For example, without an SGP (TAM), 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 Reading scale. For example, a high-achieving second-grader grows at a different rate than a low-achieving second-grader. Similarly, a high-achieving second-grader grows at a different rate than a high-achieving eighth-grader.

SGP (TAM)s 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 SGP (TAM)s 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 (TAM) of 62, that particular group of students, on average, achieved higher growth than their academic peers.

Lexile® Measures

In cooperation with MetaMetrics®, beginning in August 2014, users of STAR Reading will have the option of including Lexile measures and Lexile ZPD ranges on certain STAR Reading score reports. Reported Lexile measures will range from BR400L to 1825L. (The "L" suffix identified the score as a Lexile measure. Where it appears, the "BR" prefix indicates a score that is below 0 on the Lexile scale; such scores are typical of beginning readers.)

Lexile ZPD Ranges

A Lexile ZPD range is a student's ZPD Range converted to MetaMetrics' Lexile scale of the readability of text. When a STAR Reading user opts to report student reading abilities in the Lexile metric, the ZPD range will also be reported in that same metric. The reported Lexile ZPD ranges are equivalent to the grade level ZPD ranges used in STAR Reading and Accelerated Reader, expressed on the Lexile scale instead of as ATOS reading grade levels.

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

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

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

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

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

Renaissance Learning's (RLI) readability measure is known as the Advantage/TASA Open Standard for Readability (ATOS). The *ATOS for Text* readability formula was developed through extensive research by RLI in conjunction with Touchstone Applied Science Associates, Inc. (TASA), now called Questar Assessment, Inc. A great many school libraries use ATOS book levels to index readability of their books. ATOS book levels, which are derived from *ATOS for Books* measures, express readability as grade levels; for example, an ATOS readability measure of 4.2 means that the book is at a difficulty level appropriate for students reading at a level typical of students in the 4th grade, 2nd month. To match students to books at an appropriate level, the widely used Accelerated Reader system uses ATOS measures of readability and student's Grade Equivalent (GE) scores on standardized reading tests such as STAR Reading.

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

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

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

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

Special STAR Reading Scores

Most of the scores provided by STAR Reading software are common measures of reading performance. STAR Reading software also determines two additional scores. They are the Zone of Proximal Development and the diagnostic code.

Zone of Proximal Development (ZPD)

The Zone of Proximal Development (ZPD) defines the readability range from which students should be selecting books in order to ensure sufficient comprehension and therefore achieve optimal growth in reading skills without experiencing frustration. STAR Reading software uses Grade Equivalents to derive a student's ZPD score. Specifically, it relates the Grade Equivalent estimate of a student's reading ability with the range of most appropriate readability levels to use for reading practice. Table 61 on page 164 shows the relationship between GEs and ZPD scores.

The Zone of Proximal Development is especially useful for students who use Accelerated Reader, which provides readability levels on over 80,000 trade books, magazines, and textbooks. Renaissance Learning developed the ZPD ranges according to Vygotskian theory, based on an analysis of Accelerated Reader book reading data from 80,000 students in the 1996–1997 school year. More information is available in *The research foundation for Accelerated Reader goal-setting practices* (2006), which is published by Renaissance Learning (http://doc.renlearn.com/KMNet/R001438603GC81D6.pdf).

Diagnostic Codes

Diagnostic codes represent general behavioral characteristics of readers at particular stages of development. They are based on a student's Grade Equivalent and Percentile Rank achieved on a STAR Reading test. The diagnostic codes do not appear on the STAR Reading Diagnostic Report, but the descriptive text associated with each diagnostic code is available on the report.

Table 54 shows the relationship between the GE and PR scores and the resulting STAR Reading diagnostic codes. Note that the diagnostic codes ending in "B" contain additional prescriptive information to better assist those students performing at or below the 25th percentile.

Table 54: Diagnostic Code Values by Percentile Rank

	Diagnostic Code			Diagnostic Code	
Grade	**PR > 25**	**PR <= 25**	**Grade**	**PR > 25**	**PR <= 25**
0.0–0.9	01A	01B	4.8–5.7	06A	06B
1.0–1.7	02A	02B	5.8–6.7	07A	07B
1.8–2.7	03A	03B	6.8–8.7	08A	08B
2.8–3.7	04A	04B	8.8–13.0	09A	09B
3.8–4.7	05A	05B			

Expert consultants from both academia and public education developed and reviewed the diagnostic codes and accompanying text using standard scope and sequence paradigms from the field of reading education. The reviewers found:

1. The diagnostic information succinctly characterizes readers at each stage of development and across grade levels 1–12;
2. Critical reading behaviors are listed for successful students at each stage of development; and
3. Corrective procedures are recommended at each stage of development that adequately address important interventions.

Grade Placement

It is very important that STAR Reading software uses the correct grade placement values when determining the norm-referenced scores. The values of PR and NCE are based not only on what scaled score the student achieved but also on the grade placement of the student at the time of the test (for example, a second-grader in the seventh month with a scaled score of 395 would have a PR of 65, while a third-grader in the seventh month with the same scaled score would have a PR of 41). Thus, it is crucial that student records indicate the proper grade when students take a STAR Reading test, and that any testing in July or August reflects the proper understanding of how STAR Reading software deals with these months in determining grade placement.

Indicating the Appropriate Grade Placement

The numeric representation of a student's grade placement is based on the specific month and day in which he or she takes a test. Although teachers indicate a student’s grade level using whole numbers, STAR Reading software automatically adds fractional increments to that grade level based on the month and day of the test. To determine the appropriate increment, STAR Reading software considers the standard school year to run from September–June and assigns increment values of .0–.9 to these months. Table 53 on page 122 summarizes the increment values assigned to each month.

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

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

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

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

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

Instructions for specifying school years and grade assignments can be found in the *Renaissance Place Software Manual*.

Compensating for Incorrect Grade Placements

Teachers cannot make retroactive corrections to a student's grade placement by editing the grade assignments in a student's record or by adjusting the increments for the summer months after students have tested. In other words, STAR Reading software cannot go back in time and correct scores resulting from erroneous grade placement information. Thus, it is extremely important for the test administrator to make sure that the proper grade placement procedures are being followed. If a student has tested with an incorrect grade placement assignment (the Growth, Screening, Summary, and Test Record Reports include grade placement), the procedures outlined on page 126 in the discussion about Table 58 can be used to arrive at corrected estimates for the student's Percentile Rank and Normal Curve Equivalent scores.

STAR Reading in the Classroom

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

Goal Setting and Instructional Planning

Goal setting is an almost ubiquitous practice in education. Teachers continually set goals for their students and administrators set goals for their schools. By setting clear and achievable goals people are able to comport their behavior in an appropriate manner towards achieving those goals. This is true of school-wide or classroom-specific goals. However, not all goals are set equally. Some goals may be set ambiguously or lack a clear and measurable frame of reference. Good goal setting includes setting realistic and measurable goals that are achievable within the time frame identified.

Goals can provide a clear set of expectations of what must be accomplished and in what amount of time. It is also possible to break down long-term goals into a series of intermediate objectives or short-term goals. This can help to focus time and energy on the important aspects of meeting the long-term goal at shorter and more manageable increments. It also provides a standard for which a person may strive. Goals can also be motivating in that the realization of them provides a sense of accomplishment and achievement.

There are a few crucial aspects of goal setting in general. One of the essential aspects of goal setting is to set a measurable goal objective for some point in the future. This goal must be measurable so as to establish a criterion that represents accomplishment. It is also useful to set a series of intermediate, measurable steps to accomplish that goal. This provides a method of incremental evaluation of the progress being made towards the long-term goal. The power of this method is that it can provide early warning signals with respect to potential problems meeting the goal or recognition that one is on-track to meeting the stated objective in the future. These types of signals are important for an objective evaluation of progress. This is one of the main reasons educators need reliable and valid measurements.

If we are to measure progress and goal attainment, we need to be sure that the measuring device actually measures what we think it measures, and that it does so consistently. As an extreme example, if our long-term goal was to have our students improve their Instructional Reading Level (IRL) and we used a math test to measure progress, we should not be surprised when the signals we receive from a math test provide no relevant information on improvements in IRL scores.

It is also important that the assessment measure we use provides consistent scores because we would like to be confident that the score a student received actually tells us with a high level of precision what the student's actual ability level is.

STAR Reading provides a reliable and valid method for measuring progress towards achievable goals in reading comprehension. By using STAR Reading on a regular basis, such as quarterly or monthly, 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 Reading also provides practical advantages over other methods of gathering multiple pieces of data over time needed for monitoring achievement towards a set goal. It takes ten minutes or less to administer; this brief administration time helps maximize the amount of in-class time available for instruction. Results are also provided immediately to the teacher so the teacher will be able to review the student's progress more quickly than with most assessments.

STAR Reading can also be administered at different times for different students and at different frequencies. This allows the teacher to specify and make professional decisions concerning intermediate assessments on a student-by-student basis. It also allows the teacher to measure a student's specific response to any type of intervention being provided. This helps to strengthen the teacher's ability to make real-time, professional decisions about instructional approaches for each student.

STAR Reading can also be administered quite frequently. This allows the results of the assessment to be graphed in order to show growth. Charting progress in this way can be used both at the individual and classroom level as an evaluative check to monitor effectiveness. Periodic charting of progress can also be motivating, as students visualize their progress and recognize their achievement. This type of ongoing information gathering can be used for a variety of different functions within a school; examples include parent-teacher meetings and child-study team meetings where groups of teachers discuss ways to intervene with struggling students.

The STAR Reading assessment also has been shown to be highly related to state assessment and widely used standardized tests. This can facilitate critical benchmarking of student achievement across the grades. STAR Reading does not specifically measure states' instructional standards, but scores on STAR Reading assessments are statistically related to those proficiency standards. Therefore, scores on STAR Reading can be used to predict later outcomes. This type of information is useful in forecasting educational achievement and making decisions about utilizing resources with respect to a student's instruction. It is also possible to employ more complex, school- or district-wide implementations of the assessment to gauge student progress towards the all-important end-of-year goals consistent with a state's educational standards.

To interpret screening results, schools often use benchmarks and cut scores. These scores help educators identify which students require some form of intervention to accelerate growth and move toward proficiency. A goal-setting wizard is used in the program to set and track goals; the Screening Report and the Student Progress Monitoring Report are used to track students' progress towards goals and growth. (See the *STAR Reading Software Manual* for more information.)

Formative Assessment

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

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

- Provide an accurate estimate of students' current reading level so teachers can match students with appropriate texts for recreational and content-area reading
- Ensure that students are reading more difficult books as their abilities increase
- Identify end-of-year goals for text difficulty
- Help students choose books from different genres that match their interests and challenge their abilities

Measuring Growth

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

Absolute versus Relative Growth

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

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

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

As an example, imagine a group of students whose test results place them at the 40th percentile, with an average Scaled Score of 519, in the fall of grade 5. In the fall of grade 6, the same group still scores at the 40th percentile with an average Scaled Score of 611. This group of students has experienced 92 Scaled Score points of absolute growth, but there has been no relative growth (since the group scored at the 40th percentile in both grade distributions). In other words, relative growth will only be positive when growth has exceeded "normal" growth as defined by the norming or reference sample. In general, norm-referenced scores such as percentiles and Normal Curve Equivalent scores only indicate relative growth, whereas Scaled Scores (and Grade Equivalent scores) reflect absolute growth. The STAR Reading Growth Report provides you with information about both aspects of growth. In general, most educational program evaluation designs attempt to determine whether relative growth has occurred. That is, they are attempting to measure the impact of the intervention or program, above and beyond normal growth.

Methods of Measuring Growth

New interventions are continually being proposed for educational settings, most with the aim of improving educational outcomes. Such interventions may be extensive, such as a new teaching method or new curriculum, or they may be smaller in scope, such as a new textbook. The introduction of a Tier 1 progress-monitoring system, such as Accelerated Reader, into a school or classroom is a good example of such an intervention. Whatever the proposed intervention, however, it is first necessary to establish its effectiveness in terms of the educational benefit for students. Examination of the effectiveness of new teaching methods, a new curriculum, and other such interventions is extremely important if we are to accurately determine whether these programs and/or methods are working. This is important for appropriate direction of limited resources and for ensuring that those programs, which will have the most educational impact on children, are clearly identified.

Along with identifying whether or not an intervention is effective by use of a final summative evaluation, ongoing formative evaluations are also important. The evaluation of student progress is an ongoing procedure as the students learn and apply principles and facts learned in the classroom to solve everyday problems. Therefore, the measurement of growth can be seen as a descriptive method for understanding the developmental path of students as they acquire certain skills and enhance other abilities. With the use of ongoing monitoring of progress, teachers may be able to intervene more quickly to alter the course of instruction for a group or even more specifically to an intervention targeted at one or a few students who may be struggling. However, the monitoring of progress on an individual or small group basis is not limited to only students with high needs, but can also be used to monitor the progress of high-achieving students who may be provided more free time to explore individual interests.

The measurement of growth is a long-established tradition in social sciences in general and education specifically. While this is a large and important area of exploration, the depth of methodological and statistical analysis available at present cannot be fully described in a technical manual. The intention of the following sections is to provide a general overview of possible methods of evaluating growth using STAR Reading. We also provide a set of reference material at the close of the *Technical Manual* for interested readers to pursue a more thorough investigation of current methods of analysis and design (see page 171).

Pretest/Posttest Designs

One of the simplest methods for evaluating the effect of an intervention is the pretest-posttest paradigm, in which students are assessed twice—once prior to intervention, and once again at its completion. This method was born out of the experimental methodologies of science in an effort to quantify changes in an outcome variable by isolating the independent variables in a given system. For instance, if one would like to know if a specific intervention increases multiplication skills or phonemic segmentation, one would isolate a sample of students, randomly assign half of the students to a no-intervention group and the other half to intervention, and assess all of them before and after the intervention. Then one would look for differences in outcomes between the two groups, assuming the intervention is the only systematic difference between the groups, and make a claim about whether or not the students in the intervention group did better when compared to the students who did not receive the intervention (the no-intervention model).

An experiment with a pretest/posttest design can utilize a control group of students, who, like the above example, do not receive the intervention. This provides a comparison group against which to gauge the practical effects of the intervention applied to the intervention or treatment group and make inferences about intervention effectiveness over and above those without the intervention.

However, sometimes the use of a control group is not feasible. Under these circumstances, educators may opt to utilize norm-referenced scores, such as Percentile Ranks or Normal Curve Equivalent (NCE) scores. For example, a school may introduce a new curriculum to a whole grade level and thus would not have a readily available control group. The school may decide to use a “proxy” comparison group by utilizing norm-referenced scores. In effect, the test developer’s norming group is being used as a proxy for a control group who are not provided the intervention. This allows for changes in relative growth to be evaluated against the norming group.

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 whether 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 technical adequacy of the test itself (e.g., 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 form 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 tests with no items in common should be administered; STAR Reading is ideal for this, because tests administered to a student within 180 days of one another will have no items in common.

When the test scores used in the evaluation are norm-referenced (such as Percentile Ranks), then a control group is not necessarily required since the scores themselves allow you to measure growth relative to the peer (norming) group. It should be noted that when a test is normed, the percentile information is derived based on the specific point during the academic year when the test was administered. For example, suppose that a test was normed in the spring (seven months into the school year), but a teacher wants to make an assessment at the beginning of the school year. In order to provide normative information for each month of the academic year, STAR Reading software examines the difference between adjacent grade levels and, presuming even growth, interpolates between the empirical (observed) norms. Caution should be exercised when looking at growth that is based on these interpolated percentiles. This is because the assumption that growth occurs evenly over the time period (i.e., between the adjacent empirical percentiles) may be unrealistic.

The goal of this type of study is to determine whether a program intervention has resulted in improvement beyond what is expected based on the norming population (i.e., to see if the posttest results place the students above where they would be if there had not been any intervention). For example, if a group of 4th-grade students' pretest scores indicate that their group percentile (corresponding to the average NCE) is 25, then we want to see whether their 5th-grade posttest scores will result in a group percentile that is greater than 25. Caution must be exercised in cases where average pretest scores are substantially above or below the norm, however. Due to the phenomenon known as "regression to the mean," posttest scores will tend to move towards the norms group mean even if no real change has occurred. Consequently, corrections for regression to the mean may need to be applied before the results of an experimental intervention are interpreted.

When comparing the students' growth to growth based on norms, only one group is required, but in this case, the time period between pretest and posttest should be at least one year; otherwise the growth would be referenced against interpolated data. This corresponds with US Department of Education recommendations for Chapter I (Title I) program impact studies, which state that:

> The general rule of thumb for norm-referenced evaluations is that testing should be done within two weeks of the midpoint of the empirical norming period (U.S.D.E. Evaluator's References for Title I Evaluation and Reporting System, Volume 2).

For the STAR Reading 2 test, the empirical norming period was in the month of April. The US Department of Education further recommends that interpolated norms that vary by more than six weeks from the empirical data points should not be used for norm-referenced evaluations. In general, a good rule of thumb regarding sample size requirements for any growth study is “more is better.” As the size of the group increases, you can be more confident that the obtained results are genuine.

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

Longitudinal Designs

Longitudinal designs are different from pretest/posttest designs in that data is gathered on the same students multiple times over an extended time period. Some people argue that the evaluation of only two time points like the pretest/posttest design does not successfully identify a longitudinal design. A longitudinal design has at least three time points of measurement. An example of this approach can be seen in the assessment of students in the fall, winter, and spring quarters of the school year.

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

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

There are three highly relevant uses of longitudinal data in education. They are for use in progress monitoring, evaluating response to interventions and periodic improvement estimates. These will be discussed in the following subsections.

Student Growth Percentile (Time-Adjusted Model) (SGP (TAM))

Because STAR Reading 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 Reading reporting in 2008 via decile-based growth norms. SGP (TAM)s represent the latest advancement in helping educators understand student growth. SGP (TAM)s are available in STAR Reading for grades 1–12.

SGP (TAM)s are a normative quantification of individual student growth derived using quantile regression techniques. An SGP (TAM) compares a student's growth to that of his or her academic peers nationwide. SGP (TAM)s from STAR Reading provide a measure of how a student changed from one STAR testing window[5] to the next, relative to other students with similar starting STAR Reading scores. SGP (TAM)s 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 (TAM) 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 Reading score.

In applying the SGP (TAM) approach to STAR data, Renaissance Learning has worked closely with the lead developer of SGP (TAM), Dr. Damian Betebenner, of the Center for 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 (TAM) was initially developed for measuring growth on state tests across years, applying the SGP (TAM) approach to interim, within-year assessment data involved a number of technical challenges, primarily the differences regarding how STAR Reading 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 Reading 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 Reading 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 Reading on a more frequent basis to inform instructional decisions, such as whether the student is responding adequately to an intervention.

[5] 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.

Because of this flexibility, not all students necessarily take STAR Reading at the same time; the number and dates of administration may vary from one student to the next. However, the majority of students test within at least two of the following time periods during the school year: fall (August 1–November 30), winter (December 1–March 31), and/or spring (April 1–July 31). We chose these date ranges when defining the data sets that would be used to determine Student Growth Percentiles. Therefore, we can provide Student Growth Percentiles for achievement that takes place between fall and winter STAR Reading testing, winter and spring STAR Reading testing, and/or fall and spring STAR Reading testing, as defined above.

To calculate Student Growth Percentiles, Renaissance Learning collected hosted student data from the two most recent school years (2011–12 and 2012–13). Table 55 has details on the demographics of these students. 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 (TAM) 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 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 cutpoints for the weekly growth rate conditional achievement distribution associated with the given initial score. These cutpoints 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 Reading 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 cutpoints, we are able to calculate a Student Growth Percentile for every subject, grade, and score combination.

Table 55: Sample Characteristics, STAR Reading SGP (TAM) Study

		Sample %		
		Fall to Spring (n = 3,528,829)	**Fall to Winter (n = 4,019,291)**	**Winter to Spring (n = 4,114,791)**
Geographic Region	Midwest	22.1%	21.0%	22.2%
	Northeast	9.7%	8.7%	9.7%
	South	47.9%	49.9%	48.8%
	West	20.3%	20.4%	19.3%
	Response Rate	97.8%	97.6%	97.7%
School Type	Public	96.2%	96.1%	96.4%
	Private, Catholic	2.4%	2.5%	2.3%
	Private, Other	1.4%	1.4%	1.3%
	Response Rate	93.6%	93.3%	93.3%
School Enrollment	< 200	3.4%	3.5%	3.5%
	200–499	36.5%	36.6%	36.9%
	500–2,499	59.8%	59.6%	59.4%
	2,500+	0.2%	0.3%	0.2%
	Response Rate	95.1%	94.8%	94.9%
School Location	Urban	28.2%	28.3%	27.9%
	Suburban	27.5%	27.1%	27.8%
	Town	16.1%	16.4%	16.1%
	Rural	28.2%	28.2%	28.1%
	Response Rate	89.1%	88.7%	89.0%

Table 55: Sample Characteristics, STAR Reading SGP (TAM) Study (Continued)

		Sample %		
		Fall to Spring (n = 3,528,829)	**Fall to Winter (n = 4,019,291)**	**Winter to Spring (n = 4,114,791)**
Ethnic Group	Asian	3.6%	3.5%	3.6%
	Black	25.1%	26.3%	24.8%
	Hispanic	18.3%	18.1%	18.7%
	Native American	1.7%	1.8%	1.9%
	White	51.4%	50.3%	51.1%
	Response Rate	45.6%	44.0%	44.5%
Gender	Female	49.1%	49.0%	49.0%
	Male	50.9%	51.0%	51.0%
	Response Rate	78.8%	77.6%	78.3%

Periodic Improvement

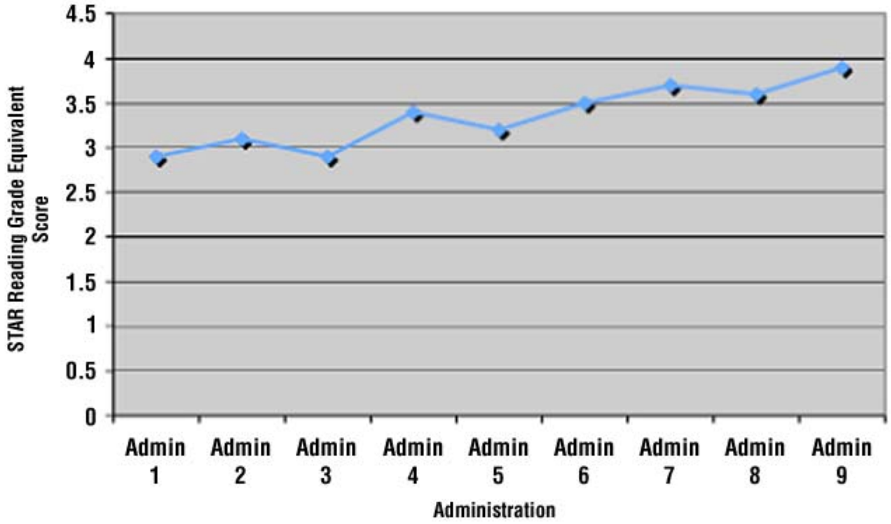
The Grade Equivalent Score can be used for measuring periodic improvement because it is reported in tenths of a grade. The correspondence between decimal value and month is shown in Table 56.

Table 56: Correspondence between Decimal Value and Month

Month	Decimal Equivalent	Month	Decimal Equivalent
September	0.0	February	0.5
October	0.1	March	0.6
November	0.2	April	0.7
December	0.3	May	0.8
January	0.4	June	0.9

The Grade Equivalent score generated by STAR Reading makes it possible to track the progress students should make on a monthly and annual basis. It is important to keep in mind, however, that the month-to-month Grade Equivalent Scores for a student are unlikely to move upward consistently. Students making appropriate progress may nonetheless show an erratic growth trajectory. Figure 7 shows the score trajectory for a typical third-grade student for nine monthly administrations of STAR Reading.

Figure 7: Monthly Progress of a Third Grader



The student started the year a little below the 3.0 GE at approximately a GE of 2.9 and is showing approximately a year's growth 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 Grade Equivalent Score from one test to the next is not a matter of concern unless it persists for two or more assessments. Intermittent score declines and erratic trajectories are not unique to STAR Reading. 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, 2003).

STAR Reading provides an efficient and useful measure of growth for both formative and summative evaluations using both pretest/posttest and longitudinal designs. STAR Reading addresses many of the problems normally associated with measuring growth over time. One of those is the time involved in assessing multiple students many times throughout the year. With STAR Reading, each student can take the assessment in about 10 minutes and at any time during the monthly period. Therefore, using STAR Reading, the teacher can maximize instructional time for the class as a whole and minimize the assessment time for each student. Also, since the scoring is done automatically, the teacher is able to receive rapid feedback without the time associated with scoring each student's assessment protocol.

In the context of progress monitoring, RTI and periodic improvement methods, STAR Reading provides a reliable and valid, norm-referenced measure of a student's reading ability. This can be used to establish a baseline measure of student ability and to evaluate student growth over time. This type of information is vital since many times in the educational setting one is unable to define a control or reference group to which one will make later comparisons.

Growth Estimates

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

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

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

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

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

To estimate the normal amount of growth from year to year, one must take into account both the grade level of the student at the time of the initial evaluation and also the performance level of the student. To facilitate the use of STAR Reading scores for estimating growth for students one can use the normative data or one can use empirical data derived from one's own district or school. The use of empirical support for making estimates about growth will be developed in the following section with examples.

Progress Monitoring

Beginning with the STAR Reading version 4.3, Renaissance Place editions include Annual Progress reports. Each of these reports contains graphical displays of individual and class scores that include STAR Reading scores from all tests administered within the current school year. Using this report, teachers can compare student progress with that of a national normed group of students in the same grade.

Research Support

A number of research projects used STAR Reading to help teachers plan instruction. A study by Borman and Dowling (2004) of the University of Wisconsin found that information provided by STAR Reading contributed to improved teacher planning and student achievement. Teachers used STAR Reading to match students with appropriate books as part of a study of the effectiveness of reading practice.

> Also consistent with the RR [Reading Renaissance] program theory, the student level results suggest that a high success rate over the course of the school year predicts better outcomes at the end of the year. This finding is consistent across all samples, the elementary, middle-school, and high-school groups. In contrast to the general theory of the model, though, after controlling for students' baseline scores, number of words read, and reading success rate, students who were assigned reading material that was, on average, beyond their baseline ability performed better on the posttest than did students who were assigned material within their optimum reading range. Consistent with the theory, though, students who were assigned material below their optimum reading range performed worse on the outcome than did students who read material that tended to be within the optimum range. This result suggests that if students' success rates are not suffering, teachers should modify their plans and assign material to students that is above their apparent baseline ability. In this respect, the finding supports suggestions provided to teachers by RR to adjust book levels if the suggested optimum range appears to be too easy or difficult for the student. This result was relatively uniform across all three groups—elementary, middle school, and high school—we studied. (pp. 25–26)

STAR Reading was also used as a planning and assessment tool in a study conducted by Sadusky and Brem (2002). Scores on the SAT9 and STAR Reading were highly correlated (between 0.65–0.75), and STAR Reading was used to develop a unique approach to having students select books that are consistent with their reading abilities and their interests. In addition, the Reading Renaissance model, of which STAR Reading is an important component, proved to be motivating and a critical planning tool.

> In the Reading Renaissance model there are three levels of student goals that are set based upon each student's initial STAR test results. The program provides guidance for teachers to then set point and level goals for each student. The second tier of goal setting involves the classroom as a group. Following prescribed calculations, teachers are able to determine yearly point and level goals for the classroom. When classrooms attain a Model Classroom banner you can hear the group cheering gleefully, and it doesn't take long for that banner to be displayed proudly outside the classroom door. (p. 28)

The Center for Research in Education Policy of the University of Memphis conducted a study of the School Renaissance model. The report entitled "The Effect of School Renaissance on Student Achievement in Two Mississippi School

Districts” (Ross & Nunnery, 2005) is available on the University’s web site (http://www.eric.ed.gov/PDFS/ED484275.pdf). The researchers concluded the following about STAR Reading:

> Positive aspects included the diagnostic component (STAR testing) and the ease with which students were assigned to their appropriate reading level. (p. 2)

In a second study reported on the website, STAR Reading was the progress measurement in a randomized experiment testing the effects of the Reading Renaissance model in an urban school district.

Holmes and Brown at the University of Georgia used STAR Reading to evaluate the effectiveness of the School Renaissance model in Georgia schools. The researchers stated that:

> …this study sought to follow a cohort of children across three grades to evaluate the effects of implementation of School Renaissance on the progress of individual children…In all nine comparisons involving scores in reading, language arts, and mathematics, the Renaissance schools’ children outperformed the contrast schools’ children. (p. 21)

In an independent study of test scores from 1,100 predominantly Hispanic students in grades 3–6, Bennicoff-Nan sought to determine the predictive ability of STAR Reading for high-stakes assessments that were part of the accountability system in California, including the SAT9 and the California Standards Test (CST) for English/Language Arts. Moderately strong to very strong correlations were found between STAR Reading and these tests across all grades analyzed; correlation coefficients ranged from 0.69–0.87. The author concludes that STAR Reading is an efficient use of time and labor in monitoring student progress in reading in the classroom, and recommends its use by California school administrators to measure progress toward state accountability goals.

STAR Reading and No Child Left Behind

STAR Reading may be useful for districts and schools as they conform to the 2001 No Child Left Behind Act. For example, the No Child Left Behind Act requires states, starting in 2005, to annually measure the reading progress of students in grades 3–8. As noted throughout this manual, STAR Reading is a reliable and valid measure of reading achievement for students in grade K–12. Furthermore, due to its computer-adaptive features, STAR Reading requires less administration time and supervision than paper and pencil tests without compromising the psychometric quality of scores.

No Child Left Behind also requires that federal funding go only to those reading programs that are backed by scientific evidence. As noted in the above section on growth measurement, teachers and administrators can use STAR Reading to evaluate the effectiveness of reading programs and interventions. Given the increased emphasis being placed on using only research-based teaching methods, more and more teachers will find STAR Reading an invaluable tool in the process of demonstrating growth in reading achievement resulting from their reading programs.

Conversion Tables

Table 57: Scaled Score to Grade Equivalent Conversions

SS Range		
Low	**High**	**Grade Equivalent**
0	1	0
7	8	0.1
9	15	0.2
20	21	0.3
22	28	0.4
29	35	0.5
36	42	0.6
43	49	0.7
50	55	0.8
56	62	0.9
63	68	1
69	73	1.1
74	81	1.2
82	92	1.3
93	105	1.4
106	120	1.5
121	137	1.6
138	153	1.7
154	171	1.8
172	188	1.9
189	206	2
207	223	2.1
224	240	2.2
241	257	2.3
258	273	2.4

Table 57: Scaled Score to Grade Equivalent Conversions (Continued)

SS Range		
Low	**High**	**Grade Equivalent**
274	288	2.5
289	303	2.6
304	317	2.7
318	330	2.8
331	343	2.9
344	355	3
356	367	3.1
368	378	3.2
379	389	3.3
390	399	3.4
400	409	3.5
410	419	3.6
420	428	3.7
429	437	3.8
438	446	3.9
447	455	4
456	465	4.1
466	474	4.2
475	483	4.3
484	492	4.4
493	502	4.5
503	511	4.6
512	521	4.7
522	531	4.8
532	542	4.9
543	552	5
553	563	5.1
564	574	5.2

Table 57: Scaled Score to Grade Equivalent Conversions (Continued)

SS Range		
Low	High	Grade Equivalent
575	586	5.3
587	597	5.4
598	609	5.5
610	621	5.6
622	633	5.7
634	645	5.8
646	658	5.9
659	671	6
672	683	6.1
684	696	6.2
697	709	6.3
710	722	6.4
723	735	6.5
736	748	6.6
749	761	6.7
762	774	6.8
775	787	6.9
788	799	7
800	812	7.1
813	824	7.2
825	836	7.3
837	848	7.4
849	860	7.5
861	872	7.6
873	884	7.7
885	895	7.8
896	906	7.9
907	917	8

Table 57: Scaled Score to Grade Equivalent Conversions (Continued)

SS Range		Grade Equivalent
Low	High	
918	928	8.1
929	938	8.2
939	949	8.3
950	959	8.4
960	969	8.5
970	978	8.6
979	988	8.7
989	998	8.8
999	1007	8.9
1008	1016	9
1017	1025	9.1
1026	1034	9.2
1035	1043	9.3
1044	1052	9.4
1053	1060	9.5
1061	1068	9.6
1069	1077	9.7
1078	1085	9.8
1086	1093	9.9
1094	1101	10
1102	1109	10.1
1110	1116	10.2
1117	1124	10.3
1125	1131	10.4
1132	1138	10.5
1139	1145	10.6
1146	1151	10.7
1152	1157	10.8

Table 57: Scaled Score to Grade Equivalent Conversions (Continued)

SS Range		Grade Equivalent
Low	High	
1158	1163	10.9
1164	1169	11
1170	1175	11.1
1176	1180	11.2
1181	1184	11.3
1185	1189	11.4
1190	1193	11.5
1194	1197	11.6
1198	1200	11.7
1201	1203	11.8
1204	1207	11.9
1208	1210	12
1211	1213	12.1
1214	1216	12.2
1217	1219	12.3
1220	1223	12.4
1224	1228	12.5
1229	1233	12.6
1234	1240	12.7
1241	1248	12.8
1249	1258	12.9
1259	1400	13

Table 58: Scaled Score to Percentile Rank Conversions[a]

	Grade Placement											
PR	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**	**6.0**	**7.0**	**8.0**	**9.0**	**10.0**	**11.0**	**12.0**
1	57	72	88	141	198	265	301	340	391	418	418	418
2	59	75	97	170	229	298	341	379	436	461	461	475
3	60	77	106	192	253	324	367	410	463	498	498	521
4	61	78	122	210	272	345	388	436	488	526	526	551
5	61	80	138	224	287	362	408	453	515	554	554	575
6	62	82	150	235	301	374	427	467	530	575	579	607
7	63	83	160	246	314	389	443	485	550	599	604	627
8	63	85	169	256	325	400	454	501	564	620	627	654
9	64	86	176	264	336	413	463	516	579	636	642	674
10	64	87	184	271	345	424	471	527	594	648	657	692
11	65	88	190	276	355	434	481	541	607	661	672	711
12	65	89	196	282	363	444	491	554	621	674	684	725
13	65	91	203	288	369	452	499	562	633	687	703	745
14	66	93	208	294	374	458	508	570	644	703	718	768
15	66	96	214	301	382	465	516	580	656	716	735	788
16	67	98	219	307	390	471	523	589	665	728	758	809
17	67	100	225	313	396	478	532	601	676	743	776	837
18	67	102	229	318	402	486	541	610	687	761	788	851
19	67	105	234	323	410	493	551	620	698	776	801	864
20	68	107	239	329	417	499	556	628	710	788	816	881
21	68	110	243	334	424	506	562	636	723	802	831	892
22	68	115	248	339	430	513	569	646	736	817	843	900
23	69	120	253	344	437	518	576	655	755	832	853	909
24	69	125	258	349	443	524	584	664	772	843	865	916
25	69	131	263	354	448	531	591	673	783	852	879	925
26	70	136	267	359	452	538	599	681	793	861	889	939
27	70	140	271	363	456	546	607	691	803	875	898	949

Table 58: Scaled Score to Percentile Rank Conversions[a] (Continued)

	Grade Placement											
PR	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**	**6.0**	**7.0**	**8.0**	**9.0**	**10.0**	**11.0**	**12.0**
28	70	144	275	366	460	552	615	701	814	884	905	960
29	70	148	278	370	465	557	622	710	828	892	911	968
30	71	153	282	373	469	561	629	721	837	899	919	976
31	71	157	286	377	473	566	635	731	846	905	929	994
32	71	161	290	381	478	571	642	743	853	910	940	1015
33	72	164	294	387	483	577	649	759	864	916	949	1034
34	72	169	299	391	489	583	656	772	877	923	960	1050
35	72	173	303	396	494	588	664	785	885	933	968	1063
36	72	176	307	400	497	594	671	794	893	943	974	1080
37	73	180	311	404	502	602	679	804	899	952	983	1096
38	73	184	314	409	507	608	686	815	905	960	995	1108
39	73	188	318	414	513	613	695	829	910	967	1011	1121
40	73	191	322	419	517	618	704	838	916	973	1032	1135
41	74	195	325	423	520	624	712	847	922	983	1046	1149
42	74	199	329	428	525	630	720	855	932	996	1062	1161
43	74	202	333	433	529	635	728	865	941	1015	1078	1173
44	75	206	337	438	535	640	739	875	950	1034	1098	1179
45	75	209	341	443	540	646	752	883	959	1048	1106	1187
46	75	213	344	447	546	652	767	891	967	1060	1120	1198
47	75	216	347	450	552	658	777	898	973	1071	1135	1207
48	76	220	352	454	556	664	786	902	981	1091	1149	1214
49	76	223	356	457	559	670	794	907	990	1100	1160	1218
50	76	226	359	460	564	676	803	913	1010	1109	1169	1225
51	77	229	362	464	568	682	813	920	1029	1125	1177	1231
52	77	233	365	467	572	690	826	927	1044	1142	1185	1241
53	78	237	368	471	577	698	836	937	1059	1155	1194	1250
54	78	241	371	475	583	706	844	947	1076	1164	1204	1254
55	79	244	374	479	587	714	850	956	1096	1171	1213	1260

Table 58: Scaled Score to Percentile Rank Conversions[a] (Continued)

	Grade Placement											
PR	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**	**6.0**	**7.0**	**8.0**	**9.0**	**10.0**	**11.0**	**12.0**
56	79	248	377	484	592	722	856	964	1106	1179	1219	1269
57	80	252	382	489	599	730	865	970	1124	1186	1225	1279
58	80	256	386	493	606	741	874	977	1140	1195	1232	1289
59	81	259	391	497	611	755	883	988	1153	1205	1244	1293
60	81	263	395	500	616	768	891	1004	1165	1214	1252	1297
61	82	267	399	505	621	777	897	1024	1174	1219	1257	1301
62	83	270	402	510	627	786	903	1040	1183	1225	1266	1306
63	83	274	407	514	632	795	908	1054	1193	1233	1278	1310
64	84	278	412	518	637	804	914	1067	1205	1244	1289	1313
65	85	281	417	522	643	813	920	1088	1214	1253	1294	1315
66	86	285	422	526	649	825	929	1101	1220	1262	1299	1317
67	86	289	427	530	655	835	940	1114	1227	1273	1305	1319
68	87	294	432	536	662	844	950	1130	1233	1283	1308	1321
69	88	298	438	542	668	851	961	1146	1246	1292	1313	1323
70	89	303	443	548	674	857	968	1160	1254	1296	1315	1325
71	90	308	448	554	680	869	976	1169	1261	1301	1317	1327
72	92	313	452	558	688	879	987	1179	1273	1306	1319	1328
73	95	317	456	562	697	887	1005	1188	1288	1311	1321	1329
74	97	321	460	567	706	896	1025	1201	1294	1314	1324	1331
75	99	326	465	572	715	903	1044	1211	1300	1317	1326	1333
76	102	331	469	577	724	908	1060	1216	1308	1320	1328	1335
77	105	336	474	584	735	914	1085	1225	1314	1322	1330	1336
78	108	340	480	590	751	922	1102	1233	1317	1325	1332	1339
79	116	345	487	598	768	931	1119	1248	1319	1326	1334	1341
80	125	350	493	606	780	944	1138	1253	1320	1328	1336	1342
81	134	356	499	612	792	955	1158	1264	1323	1329	1338	1342
82	142	361	506	620	803	965	1173	1280	1326	1331	1340	1343
83	151	366	513	628	817	972	1186	1292	1328	1334	1341	1344

Table 58: Scaled Score to Percentile Rank Conversions[a] (Continued)

	Grade Placement											
PR	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**	**6.0**	**7.0**	**8.0**	**9.0**	**10.0**	**11.0**	**12.0**
84	158	371	518	635	832	982	1201	1296	1331	1337	1342	1344
85	167	376	524	644	845	1003	1215	1304	1334	1340	1343	1344
86	176	384	531	654	855	1027	1227	1313	1337	1341	1344	1345
87	185	392	541	664	870	1048	1242	1317	1341	1342	1344	1345
88	194	399	553	675	887	1070	1254	1323	1342	1343	1344	1345
89	204	409	560	688	899	1097	1265	1326	1343	1344	1345	1345
90	214	419	569	704	909	1120	1282	1328	1344	1344	1345	1346
91	225	431	582	721	921	1151	1295	1329	1344	1345	1345	1346
92	238	445	592	743	940	1171	1303	1332	1344	1345	1346	1346
93	253	456	607	775	962	1194	1311	1337	1345	1345	1346	1346
94	267	467	623	800	977	1216	1316	1342	1346	1346	1346	1346
95	283	485	642	837	1022	1242	1324	1343	1346	1346	1346	1347
96	305	505	672	872	1071	1273	1331	1344	1346	1346	1347	1347
97	334	534	715	920	1149	1308	1339	1345	1346	1347	1347	1350
98	377	582	847	1018	1222	1328	1345	1346	1350	1350	1353	1363
99	1400	1400	1400	1400	1400	1400	1400	1400	1400	1400	1400	1400

[a] Each entry is the highest Scaled Score for that grade and percentile.

Table 59: Percentile Rank to Normal Curve Equivalent Conversions

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

Table 60: Normal Curve Equivalent to Percentile Rank Conversion

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

Table 61: Grade Equivalent to ZPD Conversions

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

Table 61: Grade Equivalent to ZPD Conversions (Continued)

	ZPD Range			ZPD Range			ZPD Range	
GE	Low	High	GE	Low	High	GE	Low	High
2.7	2.4	3.4	7.1	4.3	7.1	11.5	4.9	11.5
2.8	2.5	3.5	7.2	4.3	7.2	11.6	4.9	11.6
2.9	2.5	3.5	7.3	4.4	7.3	11.7	4.9	11.7
3.0	2.6	3.6	7.4	4.4	7.4	11.8	4.9	11.8
3.1	2.6	3.7	7.5	4.4	7.5	11.9	4.9	11.9
3.2	2.7	3.8	7.6	4.4	7.6	12.0	4.9	12.0
3.3	2.7	3.8	7.7	4.4	7.7	12.1	4.9	12.1
3.4	2.8	3.9	7.8	4.5	7.8	12.2	4.9	12.2
3.5	2.8	4.0	7.9	4.5	7.9	12.3	4.9	12.3
3.6	2.8	4.1	8.0	4.5	8.0	12.4	4.9	12.4
3.7	2.9	4.2	8.1	4.5	8.1	12.5	5.0	12.5
3.8	2.9	4.3	8.2	4.5	8.2	12.6	5.0	12.6
3.9	3.0	4.4	8.3	4.5	8.3	12.7	5.0	12.7
4.0	3.0	4.5	8.4	4.5	8.4	12.8	5.0	12.8
4.1	3.0	4.6	8.5	4.6	8.5	12.9	5.0	12.9
4.2	3.1	4.7	8.6	4.6	8.6	13.0	5.0	13.0
4.3	3.1	4.8	8.7	4.6	8.7			

Table 62: Scaled Score to Instructional Reading Level Conversions[a]

Low	High	IRL	Low	High	IRL	Low	High	IRL
0	124	Pre-Primer (PP)						
125	159	Primer (P)						
160	168	1.0	608	616	5.0	1,012	1,022	9.0
169	176	1.1	617	624	5.1	1,023	1,034	9.1
177	185	1.2	625	633	5.2	1,035	1,042	9.2
186	194	1.3	634	642	5.3	1,043	1,050	9.3
195	203	1.4	643	652	5.4	1,051	1,058	9.4
204	212	1.5	653	662	5.5	1,059	1,067	9.5
213	220	1.6	663	673	5.6	1,068	1,076	9.6
221	229	1.7	674	682	5.7	1,077	1,090	9.7
230	238	1.8	683	694	5.8	1,091	1,098	9.8
239	247	1.9	695	706	5.9	1,099	1,104	9.9
248	256	2.0	707	725	6.0	1,105	1,111	10.0
257	266	2.1	726	752	6.1	1,112	1,121	10.1
267	275	2.2	753	780	6.2	1,122	1,130	10.2
276	284	2.3	781	801	6.3	1,131	1,139	10.3
285	293	2.4	802	826	6.4	1,140	1,147	10.4
294	304	2.5	827	848	6.5	1,148	1,155	10.5
305	315	2.6	849	868	6.6	1,156	1,161	10.6
316	325	2.7	869	890	6.7	1,162	1,167	10.7
326	336	2.8	891	904	6.8	1,168	1,172	10.8
337	346	2.9	905	916	6.9	1,173	1,177	10.9
347	359	3.0	917	918	7.0	1,178	1,203	11.0
360	369	3.1	919	920	7.1	1,204	1,221	11.1
370	379	3.2	921	922	7.2	1,222	1,243	11.2
380	394	3.3	923	924	7.3	1,244	1,264	11.3
395	407	3.4	925	928	7.4	1,265	1,290	11.4
408	423	3.5	929	930	7.5	1,291	1,303	11.5

Table 62: Scaled Score to Instructional Reading Level Conversions[a] (Continued)

Low	High	IRL	Low	High	IRL	Low	High	IRL
424	439	3.6	931	934	7.6	1,304	1,314	11.6
440	451	3.7	935	937	7.7	1,315	1,319	11.7
452	462	3.8	938	939	7.8	1,320	1,324	11.8
463	474	3.9	940	942	7.9	1,325	1,328	11.9
475	487	4.0	943	948	8.0	1,329	1,330	12.0
488	498	4.1	949	954	8.1	1,331	1,332	12.1
499	512	4.2	955	960	8.2	1,333	1,335	12.2
513	523	4.3	961	966	8.3	1,336	1,337	12.3
524	537	4.4	967	970	8.4	1,338	1,340	12.4
538	553	4.5	971	974	8.5	1,341	1,341	12.5
554	563	4.6	975	981	8.6	1,342	1,342	12.6
564	577	4.7	982	988	8.7	1,343	1,343	12.7
578	590	4.8	989	998	8.8	1,344	1,344	12.8
591	607	4.9	999	1,011	8.9	1,345	1,345	12.9
						1,346	1,400	Post-High School (PHS)

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

Table 63: Relating STAR Early Literacy Enterprise Scores to STAR Reading Scores

STAR Early Literacy Enterprise		STAR Reading			Recommended Assessment(s)
Scale Score Range	Literacy Classification	Scale Score Range	GE	ZPD Range	
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 63: Relating STAR Early Literacy Enterprise Scores to STAR Reading Scores (Continued)

STAR Early Literacy Enterprise		STAR Reading			
Scale Score Range	Literacy Classification	Scale Score Range	GE	ZPD Range	Recommended Assessment(s)
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 63: Relating STAR Early Literacy Enterprise Scores to STAR Reading Scores (Continued)

STAR Early Literacy Enterprise		STAR Reading			
Scale Score Range	Literacy Classification	Scale Score Range	GE	ZPD Range	Recommended Assessment(s)
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	

Appendix A: Sources for Authentic Texts

Sources for the authentic text passages include the following:

Adkins, Jan. *What If You Met a Pirate?* Book Level 6.4.

Aiken, Joan. *Lady Catherine's Necklace.* Book Level 7.0.

Ake, Anne. *The Gorilla.* Book Level 9.0.

Allaby, Michael. *Deserts and Semideserts.* Book Level 7.3.

Aller, Susan. *Christopher Columbus.* Book Level 4.4.

Anthony, Piers. *Vale of the Vole.* Book Level 7.7.

Ardagh, Philip. *The Rise of the House of McNally, or, About Time Too.* Book Level 5.5.

Ashby, Ruth. *Anne Frank.* Book Level 5.1.

Bailey, Gerry. *Underwater Machines*. Book Level 6.3.

Baldwin, Carol. *Living by a River.* Book Level 4.4.

Ballantyne, R.M. *Coral Island.* Book Level 2.1.

Banks, Lynne Reid. *The Secret of the Indian.* Book Level 5.3.

Barron, Stephanie. *Jane and the Ghosts of Netley.* Book Level 7.2.

Bellairs, John. *The Letter, the Witch, and the Ring.* Book Level 4.7.

Benchley, Nathaniel. *A Ghost Named Fred.* Book Level 2.4.

Berg, Elizabeth. *Joy School.* Book Level 3.8.

Berger, Melvin. *Do Bears Sleep All Winter? Questions and Answers About Bears.* Book Level 4.6.

Blackstone, Stella. *Storytime.* Book Level 4.5.

Blackwood, Gary. *Shakespeare's Spy.* Book Level 5.7.

Blume, Judy. *Just as Long as We're Together.* Book Level 3.7.

Blume, Judy. *Tales of a Fourth Grade Nothing.* Book Level 3.3.

Bockenhour, Mark. *Our Fifty States.* Book Level 8.6.

Bond, Michael. *Paddington Takes to TV.* Book Level 6.3.

Boraas, Tracy. *Puerto Rico.* Book Level 6.1.

Bosse, Malcolm. *Deep Dream of the Rain Forest.* Book Level 7.2.

Bourseiller, Philippe. *Volcanoes: Journey to the Crater's Edge (Adapted).* Book Level 6.7.

Boyd, Candy. *Chevrolet Saturdays.* Book Level 4.1.

Brooks, Geraldine. *Year of Wonders: A Novel of the Plague.* Book Level 6.9.

Brooks, Terry. *Tanequil.* Book Level 6.7.

Brown, Don. *Our Time on the River.* Book Level 4.8.

Brown, Rita. *The Tail of the Tip-Off.* Book Level 4.6.

Brust, Beth. *The Amazing Paper Cuttings of Hans Christian Andersen.* Book Level 6.9.

Buckey, Sarah. *Gangsters at the Grand Atlantic.* Book Level 4.3.

Busch, Phyllis. *Autumn.* Book Level 4.6.

Butterfield, Moira. *Electronics.* Book Level 8.5.

Carlson, Donna. *Moby Dick: With a Discussion of Determination.* Book Level 4.6.

Carmody, Isobelle. *Night Gate.* Book Level 5.6.

Carroll, Lewis. *Alice's Adventures in Wonderland* and *Through the Looking Glass.* Book Level 7.7.

Chapman, Gary. *Mountains.* Book Level 8.0.

Ching, Jacqueline. *The Assassination of Martin Luther King Jr.* Book Level 8.1.

Clancy, Tom. *Tom Clancy's Power Plays: Shadow Watch.* Book Level 8.8.

Clark, Mary. *No Place Like Home.* Book Level 6.2.

Claybourne, Anna. *Lizards.* Book Level 6.6.

Cleary, Beverly. *Ramona the Brave.* Book Level 4.9.

Cobb, Vicki. *Junk Food.* Book Level 6.3.

Cole, Joana. *Norma Jean, Jumping Bean.* Book Level 2.2.

Conly, Jane Leslie. *Trout Summer.* Book Level 4.3.

Costick, Kathleen. *The Prince and the Pauper With a Discussion of Respect.* Book Level 4.8.

Cox, Lynne. *Swimming to Antarctica: Tales of a Long-Distance Swimmer.* Book Level 6.6.

Crane, Stephen. *Maggie, a Girl of the Streets, and Other New York Writings.* Book Level 8.2.

Cussler, Clive. *Shock Wave.* Book Level 7.3.

Cussler, Clive. *Trojan Odyssey.* Book Level 7.5.

Dahl, Roald. *The Witches.* Book Level 4.7.

Danticat, Edwidge. *The Dew Breaker.* Book Level 6.7.

Davis, Kenneth. *Don't Know Much About Dinosaurs.* Book Level 6.3.

DeClements, Barthe. *Nothing's Fair in Fifth Grade.* Book Level 3.7.

Dennis, Jeanne. *Mystery at Crestwater Camp.* Book Level 4.2.

Dewey, Jennifer. *Family Ties: Raising Wild Babies.* Book Level 6.6.

Dickens, Charles. *David Copperfield (Unabridged).* Book Level 9.5.

Doherty, Berlie. *White Peak Farm.* Book Level 5.2.

Donnelly, Shannon. *Eleanor Roosevelt.* Book Level 4.1.

Dubowski, Cathy. *A Little Princess.* Book Level 2.6.

Dunham, Montrew. *Margaret Bourke-White: Young Photographer.* Book Level 6.1.

Eliot, George. *Silas Marner.* Book Level 9.7.

Englart, Mindi. *TV Reporter.* Book Level 5.5

Estes, Eleanor. *The Alley.* Book Level 5.4.

Evanovich, Janet. *To the Nines.* Book Level 4.2.

Fine, Anne. *Step by Wicked Step.* Book Level 4.2.

Finley, Martha. *Elsie's Girlhood.* Book Level 6.7.

Fletcher, Ralph. *Fig Pudding.* Book Level 3.9.

Frances, Dorothy. *Sea Turtles: Creatures of Mystery.* Book Level 2.9.

Galarza, Ernesto. *Barrio Boy.* Book Level 6.8.

Gallant, Roy. *Sand on the Move: The Story of Dunes.* Book Level 6.8.

George, Charles/Linda. *Ice Climbing (Sports Alive!).* Book Level 4.9.

Gerstein, Mordicai. *Behind the Couch.* Book Level 3.8.

Giff, Patricia. *Write Up a Storm With the Polk Street School.* Book Level 3.3.

Gonzalez, Catherine. *Cynthia Ann Parker, Indian Captive.* Book Level 5.7.

Goodman, Susan. *Cora Frear.* Book Level 3.5.

Gore, Wilma. *Earth Day.* Book Level 3.7.

Gourse, Leslie. *Blowing on the Changes: The Art of the Jazz Horn Players.* Book Level 8.1.

Gow, Mary. *Johnstown Flood: The Day the Dam Burst.* Book Level 5.5.

Graf, Mike. *Tornado! The Strongest Winds on Earth.* Book Level 3.5.

Graham, Pamela. *Undercover of Darkness: Animals That Move at Night.* Book Level 6.3.

Grambo, Rebecca. *Borealis: A Polar Bear Cub's First Year.* Book Level 3.7.

Gravelle, Karen. *The Driving Book: Everything New Drivers Need to Know But Don't Know to Ask.* Book Level 6.9.

Greenaway, Theresa. *Ears and Eyes.* Book Level 3.7.

Gruelle, Johnny. *Raggedy Ann & Andy: A Read-Aloud Treasury.* Book Level 5.4.

Haas, Jessie. *Fire! My Parents' Story.* Book Level 3.8.

Hacker, Carlotta. *Humanitarians.* Book Level 5.7.

Hacking, Sue. *Mount Everest and Beyond: Sir Edmund Hillary.* Book Level 4.4.

Halliday, John. *PredicKtions.* Book Level 5.6.

Harris, Tim. *Swans.* Book Level 5.5.

Hayden, Tory. *The Very Worst Thing.* Book Level 4.3.

Heinrichs, Ann. *Hawai'i (Child's World).* Book Level 3.8.

Heinrichs, Ann. *Oregon (Child's World).* Book Level 3.9.

Heinrichs, Ann. *Washington (This Land Is Your Land).* Book Level 5.4.

Herberman, Ethan. *The City Kid's Field Guide.* Book Level 6.5.

Herbert, Brian. *The Machine Crusade.* Book Level 8.2.

Higman, Anita. Pets: *Never Dance With a Tree Frog.* Book Level 3.4.

Hill, Janet. *Starlight, Star Bright.* Book Level 4.5.

Hill, Stuart. *The Cry of the Icemark.* Book Level 8.0.

Hobbs, Valerie. *How Far Would You Have Gotten If I Hadn't Called You Back?* Book Level 5.0.

Hoff, B. J. *Winds of Graystone Manor.* Book Level 7.0.

Hofmann, Ginnie. *The Runaway Teddy Bear.* Book Level 2.5.

Holub, Joan. *Charlotte's Choice.* Book Level 3.5.

Honeycutt, Natalie. *Juliet Fisher and the Foolproof Plan.* Book Level 3.6.

Hopkins, Ellen. *Tarnished Legacy: The Story of the Comstock Lode.* Book Level 4.4.

Howe, James. *What Eric Knew.* Book Level 3.9.

Huck, Charlotte. *Princess Furball.* Book Level 4.7.

Hughes, Monica. *Space Trap.* Book Level 5.1.

Hurwitz, Jane. *Choosing a Career in Animal Care.* Book Level 6.1.

Ingold, Jeanette. *Mountain Solo.* Book Level 5.3.

James, Brian. *Tomorrow, Maybe.* Book Level 4.7.

Jeffrey, Laura. *Horses: How to Choose and Care for a Horse.* Book Level 4.9.

Johnson, Charles. *Middle Passage.* Book Level 7.1.

Jordan, Shirley. *From Smoke Signals to Email.* Book Level 4.8.

Jordan, Shirley. *Pioneer Days: Moments in History.* Book Level 4.1.

Kalman, Bobbie. *Arctic Whales and Whaling.* Book Level 6.5.

Kazan, Frances. *Halide's Gift.* Book Level 6.5.

Keller, Ellen. *Animal Communication.* Book Level 4.1.

Kent, Deborah. *Dublin.* Book Level 6.3.

Kent, Deborah. *Massachusetts.* Book Level 8.7.

Kerr, Rita. *Tex's Tales.* Book Level 3.9.

Kincaid, Jamaica. *The Autobiography of My Mother.* Book Level 7.0.

Klass, David. *Danger Zone.* Book Level 5.2.

Kneale, Matthew. *English Passengers.* Book Level 7.4.

Koontz, Dean. *Lightning.* Book Level 7.5.

Kostova, Elizabeth. *The Historian: A Novel.* Book Level 7.3.

Kramer, Stephen. *Caves.* Book Level 5.6.

Krisher, Trudy. *Uncommon Faith.* Book Level 5.9.

Lackey, Mercedes. *Brightly Burning.* Book Level 7.4

Lackey, Mercedes. *Storm Warning.* Book Level 7.7.

Lackey, Mercedes. *Winds of Change.* Book Level 7.2.

Lackey, Mercedes. *Winds of Fate.* Book Level 7.3.

Landau, Elaine. *Fierce Cats.* Book Level 4.1.

Landau, Elaine. *Killer Bees.* Book Level 3.7.

Landau, Elaine. *Sinister Snakes.* Book Level 3.7.

Lauber, Patricia. *The Tiger Has a Toothache.* Book Level 3.7.

Lemieux, Jean. *Toby's Very Important Question.* Book Level 3.6.

Lenski, Lois. *Prairie School.* Book Level 4.0.

Lepthien, Emilie U. *Wetlands.* Book Level 5.5.

Lishman, Bill. *Father Goose and His Goslings.* Book Level 5.9.

Lobel, Arnold. *Frog and Toad Together.* Book Level 2.9.

London, Jack. *White Fang.* Book Level 7.4.

Lowrey, Janette. *The Poky Little Puppy.* Book Level 4.0.

Lund, Bill. *Triathlon.* Book Level 4.5.

Luttrell, Wanda. *Shadows on Stoney Creek.* Book Level 5.6.

Lutz, Norma. *William Penn: Founder of Democracy.* Book Level 6.6.

Lynch, Wayne. *Hawks.* Book Level 6.0.

Lyoie/Brissenden. *As Long as the Rivers Flow.* Book Level 3.9.

MacBride, Roger Lea. *On the Other Side of the Hill.* Book Level 5.2.

Margeson, Susan. *Viking.* Book Level 6.2.

Marrin, Albert. *Terror of the Spanish Main: Sir Henry Morgan and His Buccaneers.* Book Level 7.4.

Marsh, Carol. *The Mystery at the Boston Marathon/The Mystery on the Freedom Trail.* Book Level 5.2.

Martin, Ann. *Karen's Pilgrim.* Book Level 3.4.

Mattern, Joanne. *Fish.* Book Level 4.9.

Mayer, Mercer. *Just My Friend and Me.* Book Level 1.9.

Mazer, Harry. *The War on Villa Street.* Book Level 3.8.

McCabe, Suzanne. *Adventures of Huckleberry Finn: With a Discussion of Friendship.* Book Level 3.5.

McDaniel, Lurlene. *Lifted Up By Angels.* Book Level 4.3.

McLoone, Margo. *Women Explorers of the Oceans.* Book Level 4.5.

Meaderis, Angela. *Dare to Dream: Coretta Scott King and the Civil Rights Movement.* Book Level 6.4.

Melville, Herman. *Moby-Dick, or, The Whale.* Book Level 10.3.

Metzger, Lois. *Missing Girls.* Book Level 4.1.

Miles, Ellen. *Doctor Dolittle.* Book Level 4.9.

Miller, Sara. *Seahorses, Pipefishes, and Their Kin.* Book Level 5.8.

Miller-Schroeder, Patricia. *Bottlenose Dolphins (The Untamed World).* Book Level 6.7.

Monroe/Williamson. *First Houses: Native American Homes and Sacred Structures.* Book Level 6.8.

Morris, Deborah. *Real Kids Adventures: A Lightning Strike.* Book Level 4.1.

Morris, Deborah. *Real Kids Real Adventures: A Powerful Tornado!* Book Level 4.3.

Morris, Deborah. *Real Kids Real Adventures: Explosion!* Book Level 4.5.

Muldoon, Kathleen. *Presidential Pet "Tails."* Book Level 3.5.

Murphy, Claire. *Gold Rush Winter.* Book Level 3.1.

Nichols, Catherine. *Animal Masterminds: A Chapter Book.* Book Level 3.8.

Ninh, Bao. *The Sorrow of War: A Novel of North Vietnam.* Book Level 6.6.

Nolan, Dennis. *Wolf Child.* Book Level 5.1.

Nordhoff/Hall. *Mutiny on the Bounty.* Book Level 8.4.

O'Connor, Jane. *Sir Small and the Dragonfly.* Book Level 1.7.

Original writing by Renaissance Learning, Inc.

Orr, Tamra. *Life in the Arctic.* Book Level 5.9.

Parks, Edd Winfield. *Teddy Roosevelt: Young Rough Rider.* Book Level 3.5.

Parlin, John. *Amelia Earhart: Pioneer of the Sky.* Book Level 3.6.

Pascal/Stewart. *Jessica Plays Cupid.* Book Level 2.5.

Pascoe, Elaine. *Freshwater Fish.* Book Level 5.1.

Pascoe, Elaine. *International Space Station (Super Structures of the World).* Book Level 7.4.

Pella, Judith. *Heirs of the Motherland.* Book Level 6.9.

Perl, Lila. *Piñatas and Paper Flowers.* Book Level 6.1.

Peters, Russell. *Clambake: A Wampanoag Tradition.* Book Level 5.5.

Pevsner, Stella. *Sister of the Quints.* Book Level 3.9.

Phillips, Michael. *Wild Grows the Heather in Devon.* Book Level 7.0.

Pineiro, R.J. *Shutdown.* Book Level 7.6.

Poynter, Margaret. *Marie Curie: Discoverer of Radium.* Book Level 4.0.

Pressler, Mirjam. *Malka.* Book Level 5.9.

Price, Sean. *Robinson Crusoe: With a Discussion of Resourcefulness.* Book Level 5.1.

Price-Groff, Claire. *Thomas Alva Edison: Inventor and Entrepreneur.* Book Level 8.7.

Pupeza, Lori. Custom Bikes (Ultimate Motorcycles). Book Level 5.9.

Raskin, Ellen. *The Mysterious Disappearance of Leon (I Mean Noel).* Book Level 4.9.

Richardson, Adele. *North American Racer Snakes (Snakes).* Book Level 4.3.

Ring, Elizabeth. *Henry David Thoreau: In Step with Nature.* Book Level 5.0.

Ross, Michael. *Bird Watching With Margaret Morse Nice.* Book Level 6.4.

Sanders, Mark. *The White House.* Book Level 4.6.

Scarf, Maggi. *Meet Benjamin Franklin.* Book Level 3.3.

Scott, Caitlin. *Treasure Hunting: Looking for Lost Riches.* Book Level 3.6.

Seidler, Tor. *The Wainscott Weasel.* Book Level 4.6.

Sewell, Anna. *Black Beauty (Unabridged).* Book Level 7.7.

Sharth, Sharon. *Sea Jellies: From Corals to Jelly Fish.* Book Level 6.2.

Silate, Jennifer. *Little Sure Shot: Annie Oakley and the Buffalo Bill's Wild West Show.* Book Level 4.1.

Singer, Marilyn. *A Dog's Gotta Do What a Dog's Gotta Do: Dogs at Work.* Book Level 5.3.

Slaughter, Carolyn. *Before the Knife: Memories of an African Childhood.* Book Level 6.5.

Smith, Karla. *Virginia Plants and Animals.* Book Level 6.2.

Smith, Roland. *Thunder Cave.* Book Level 4.2.

Spalding, Andrea and David. *The Klondike Ring.* Book Level 5.0.

Spilsbury, Louise. *Why Should I Brush My Teeth? And Other Questions About Healthy Teeth.* Book Level 5.8.

Starke, Katherine. *Dogs and Puppies.* Book Level 4.1.

Staunton, Ted. *Morgan Makes a Splash.* Book Level 3.0.

Stevens, Beth. *Bicycles.* Book Level 3.0.

Stevens, Beth. *Colorful Kites.* Book Level 3.3.

Stevens, Beth. *Tops (and Other Spinning Toys).* Book Level 3.2.

Stevens, Beth. *Wheels!* Book Level 3.7.

Supples, Kevin. *Rome.* Book Level 4.6.

Takashima, Shizuye. *A Child in Prison Camp.* Book Level 3.8.

Taylor, Bonnie. *Mattie: A Brown Pelican.* Book Level 3.6.

Taylor, Bonnie. *Roscoe: A North American Moose.* Book Level 3.3.

Taylor, Bonnie. *Zelda: A Little Brown Bat.* Book Level 3.5.

Taylor, Theodore. *The Cay.* Book Level 5.3.

Temko, Florence. *Traditional Crafts from China.* Book Level 5.7.

Thackeray, William M. *Vanity Fair.* Book Level 12.4.

Thompson, Gare. *Who Was Eleanor Roosevelt?* Book Level 4.5.

Tripp, Valerie. *Felicity Learns a Lesson: A School Story.* Book Level 4.3.

Trumble, Kelly. *The Library of Alexandria.* Book Level 7.3.

Udall, Brady. *The Miracle Life of Edgar Mint.* Book Level 6.6.

Van Leeuwen, Jean. *Benjy the Football Hero.* Book Level 4.0.

Van Loon/Merriman. *The Story of Mankind.* Book Level 9.9.

VanRiper, Guernsey. *Jim Thorpe: Olympic Champion.* Book Level 4.

Verne, Jules. *20,000 Leagues Under the Sea (Unabridged).* Book Level 10.0.

Verne, Jules. *Journey to the Centre of the Earth.* Book Level 9.9.

Voltaire. *Candide.* Book Level 7.3.

Walker, Sarah. *Big Cats.* Book Level 6.2.

Wassiljewa, Tatjana. *Hostage to War: A True Story.* Book Level 4.5.

Watt, E. *Giraffes (The Untamed World).* Book Level 6.4.

Watts, Clair. *Heat Hazard: Droughts (Raintree Express).* Book Level 4.5.

Weber, William. *Care of Uncommon Pets.* Book Level 7.6.

Weil, Ann. *Volcanoes.* Book Level 4.5.

Welsbacher, Anne. *Life in a Rain Forest.* Book Level 7.4.

Wierenga, Kathy. *Croutons for Breakfast.* Book Level 4.3.

Williams, Colleen. *Homes of the Native Americans.* Book Level 6.8.

World Book Editors. *About You.* Book Level 4.8.

Appendix B: Estimated Oral Reading Fluency

Table 64: Estimated Oral Reading Fluency (Est. ORF) Given in Words Correct per Minute (WCPM) by Grade for Selected STAR Reading Scale Score Units (SR SS)

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

References

Allington, R., & McGill-Franzen, A. (2003). Use students' summer-setback months to raise minority achievement. *Education Digest, 69*(3), 19–24.

Bennicoff-Nan, L. (2002). *A correlation of computer adaptive, norm referenced, and criterion referenced achievement tests in elementary reading*. Unpublished doctoral dissertation, The Boyer Graduate School of Education, Santa Ana, CA.

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

Betebenner, D. W. (2010). New directions for student growth models. Retrieved from the National Center for the Improvement of Educational Assessment website: http://www.ksde.org/LinkClick.aspx?fileticket=UssiNoSZks8%3D&tabid=4421&mid=10564

Betebenner, D. W., & Iwaarden, A. V. (2011a). SGP: An R package for the calculation and visualization of student growth percentiles & percenitle growth trajectories [Computer Software manual]. (R package version 0.4-0.0 available at http://cran.r-project.org/web/packages/SGP/)

Betebenner, D. W. (2011b). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. The National Center for the Improvement of Educational Assessment. Retrieved from http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf

Borman, G. D. & Dowling, N. M. (2004). *Testing the Reading Renaissance program theory: A multilevel analysis of student and classroom effects on reading achievement*. University of Wisconsin-Madison.

Bracey, G. (2002). Summer loss: The phenomenon no one wants to deal with. *Phi Delta Kappan, 84*(1), 12–13.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally & Company.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.

Deno, S. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37*(3), 184–192.

Diggle, P., Heagerty, P., Liang, K., & Zeger, S. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

Duncan, T., Duncan, S., Strycker, L., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Holmes, C. T., & Brown, C. L. (2003). *A controlled evaluation of a total school improvement process, School Renaissance.* University of Georgia. Available online: http://www.eric.ed.gov/PDFS/ED474261.pdf.

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

Kirk, R. (1995). *Experimental Design: Procedures for the behavioral sciences* (3rd ed.). New York: Brooks/Cole Publishing Company.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

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

Meyer, B., & Rice, G. E. (1984). The structure of text. In P.D. Pearson (Ed.), *Handbook on reading research* (pp. 319–352). New York: Longman.

Moskowitz, D. S., & Hershberger, S. L. (Eds.). (2002). *Modeling intraindividual variability with repeated measures data: Methods and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Multiple authors. (2002). *Modeling intraindividual variability with repeated measures data: Methods and applications*. In D. S. Moskowitz & S. L. Hershberger (eds.), Mahwah, NJ: Lawrence Erlbaum Associates.

Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). New York: WCB McGraw-Hill.

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Index

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