Renaissance Place® Edition

STAR Reading®

Computer-Adaptive
Reading Test and Database

Technical Manual

Renaissance Learning
PO Box 8036
Wisconsin Rapids, WI 54495-8036
Phone: (800) 656-6740
Outside the U.S.: 1.715.424.3636
Fax: (715) 424-4242
Email: answers@renlearn.com
Support Email: support@renlearn.com
Web Site: www.renlearn.com

Copyright Notice

05/05 v.4.0

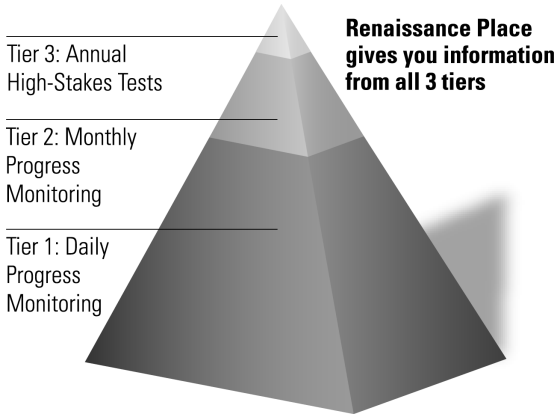
Contents

Introduction

STAR Reading and Learning Information Systems

The Renaissance Place (RP) Edition of the STAR Reading computer-adaptive test and database allows teachers to assess students' reading abilities accurately in ten minutes or less. This computer program helps educators accelerate learning and increase motivation by providing immediate, individualized feedback on student academic tasks and classroom achievement. All key decision makers throughout the district can easily access this information.

The Renaissance Place database stores all three levels of student information, including the Tier 2 data from STAR Reading:



Tier 1: Daily Progress Monitoring

Tier 1 information supports a teacher's instructional decision making, maximizes academic learning time, and ensures that the curriculum is implemented with integrity. Renaissance Tier 1 programs include Accelerated Reader, Accelerated Math, Accelerated Grammar & Spelling, Accelerated Writer, English in a Flash, Fluent Reader, and MathFacts in a Flash.

Tier 2: Monthly Progress Monitoring

Tier 2 information is gathered two to 10 times per year and is helpful in placing students at appropriate levels, measuring curriculum and instructional effectiveness, making intra-year adjustments, and predicting student and school performance on end-of-year tests. Renaissance Tier 2 programs include STAR Reading, STAR Early Literacy, STAR Math, and StandardsMaster.

Tier 3: Annual High-Stakes Testing

Tier 3 information helps measure system effectiveness. In addition to annual high-stakes tests, other sources of Tier 3 information include college entrance and advanced placement exams.

The state assessments mandated under No Child Left Behind are the most obvious examples of high-stakes tests. An effective way to ensure success on high-stakes tests is through the proper use of Tier 1 and Tier 2 programs.

STAR Reading Purpose

As a periodic progress monitoring learning information system (LIS), STAR Reading software serves two primary purposes. First, it provides educators with quick and accurate estimates of students' instructional reading levels relative to national norms. Second, it provides the means for tracking growth in a consistent manner over long time periods for all students. This is especially helpful to school- and district-level administrators.

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.x and higher RP browser-based management program provides regular, accurate reports on performance at the class, grade, building, and district level, as well as year-to-year comparisons.
- 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 district-wide. 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.

Design of STAR Reading

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 various procedures included such formats as the traditional reading passage followed by sets of literal or inferential questions; extended, previously published selections followed by open-ended questions requiring student-constructed answers; and several cloze-type procedures for passage presentation. For interrelated reasons of efficiency of assessment, objectivity and simplicity of scoring, and breadth of construct coverage, the vocabulary-in-context format was finally selected as one mode for use in assessment. For students at grade levels 1 and 2, the STAR Reading 3.x 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:

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 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 questions.

3. A child's level of vocabulary development is a major — 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 Versions 2.x and 3.x RP and higher

Since the introduction of STAR Reading Version 1.0 in 1996, STAR Reading has undergone a process of continuous research and improvement. Version 2.0 was an entirely new test, with new content and several technical innovations. Versions 3.x RP and higher are adaptations of Version 2.x designed specifically for use on a computer with web access. However, STAR Reading Versions 3.x RP and higher are identical in content to STAR Reading Version 2.x. The following improvements introduced in Version 2.0 continue to apply to Versions 3.x RP and higher.

- The item bank has been expanded from 838 test items distributed among 14 difficulty levels to 1,409 items graded into 54 difficulty levels.
- Test content has been expanded as well. STAR Reading Version 1.x consisted of a single test section that measured reading comprehension through vocabulary-in-context questions. Versions 2.x 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 has been improved. Versions 2.x 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.x 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.x and higher tests administer just 25 questions to every student. At grade levels 3 and above, there are 20 vocabulary-in-context items and five authentic text passage items. At grade levels 1 and 2, all 25 items are vocabulary-in-context items. In

contrast, Version 1.x administered a variable number of items, ranging from five to 60. The average length of Version 1.x tests was 30 items per student.

- Like the STAR Reading 1.x test before it, the STAR Reading 2.x and higher test has been nationally standardized prior to release. Therefore, its norm-referenced test scores represent the most recent benchmark available.

Improvements specific to STAR Reading versions 3.x RP and higher

In Versions 3.x 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.x to version 4.x 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.

Test Security

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

Split-application model

In the STAR 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, varied 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 a home computers). For more information on student access 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 (teacher, administrators, or staff) must enter a user name and password before they can access the data and create reports. Parents must also log in with a user name and password before they can access the Parent Report. Without an appropriate user name and password, personnel and parents cannot use the STAR 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.

Test Administration Procedures

In order to ensure consistency and comparability of results to the STAR Reading norms, students taking STAR Reading tests should follow the same administration procedures used by the norming participants. It is also a good idea to make sure that the testing environment is as free from distractions for the student as possible.

During the STAR Reading norming, the program was modified so that teachers could not deactivate the proctoring (test-monitoring) options. This was necessary to ensure that the norming data gathered were as reliable as possible. During norming, test monitors had responsibility for test security, and were required to provide access to the test for each student. In the final version of the STAR Reading test, teachers can turn off the requirement for test monitoring, but it is not recommended that they do so.

Also during norming, all of the participants received the same set of test instructions and corresponding graphics contained in the *Pretest Instructions* included with the STAR Reading product. These instructions 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 the first grade through the eighth grade. It is important to use these same instructions with all students before they take the STAR Reading test.

Test Interface

The STAR Reading test interface was designed to be both simple and effective. For purposes of standardization, the program limits input to the numeric keys found just above the letters on the standard keyboard (or on the right side of the keyboard). Students have a nearly equal footing by limiting input to only four numeric keys and the <Enter> (or <return>) key. Computer-literate students should have no advantage over those with limited computer skills.

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 STAR Reading 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.

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 3.x RP and higher test begins with items that have a difficulty level that is substantially 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 answer the first item correctly. Teachers can override this feature 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.

Once the testing session is underway, the test administers 25 items of varying difficulty based on the student's responses; this is sufficient information to obtain a reliable Scaled Score and to determine the student's Instructional Reading Level. The average length of time needed to complete a STAR Reading test is between seven and eight minutes, with a standard deviation of less than three minutes. Thus, most students will be able to complete a STAR Reading test in under ten minutes, and almost all will be able to do so in less than 13 minutes.

Test Repetition

Students can take STAR Reading tests up to five times per year at monthly intervals without concern for previous exposure to the items. The item bank supporting the STAR Reading test contains over 1400 items. STAR Reading software keeps track of the specific items presented to each student from test session to test session. By doing so, the STAR Reading software can keep item reuse to a minimum. Additionally, if a student is progressing in reading development throughout the year and from year to year, item exposure should not be an issue at all. For more information on the STAR Reading item bank's depth and breadth, see *Content and Item Development* on page 10.

Item Time Limits

The STAR Reading test has time-out limits for individual items that are based on a student's grade level. Students in first and second grade have up to 60 seconds to answer each item during their test sessions. Students in grades 3 through 12 are allowed 45 seconds to answer each vocabulary-in-context item (the first 20 items) and 90 seconds to answer each authentic text passage item (the last five test 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%) 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.

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. Therefore, students in first and second grade with the extended time limit accommodation have up to 180 seconds to answer each item. Students in grades 3 through 12 with the extended time limit accommodation have 135 seconds to answer each vocabulary-in-context item (the first 20 items) and 270 seconds to answer each authentic text passage item (the last five items).

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

Content and Item Development

Content Development

The content of STAR Reading 2.x is identical to the content in versions 3.x RP and higher. Content development was driven by the design and intended usage of the test. The desired content had to meet certain criteria. First, it had to cover a range broad enough to test students from first through twelfth grade. 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 final item bank for the STAR Reading 2.x and higher tests contains a total of 1,409 items: 1,159 vocabulary-in-context items, and 250 authentic text passage items.

The EDL Core Vocabulary

After an exhaustive search, the point of reference for developing STAR Reading items that best matched appropriate word-level placement information was found to be the Educational Development Laboratory's *A Revised Core Vocabulary* (1969). The EDL vocabulary list is a soundly developed, validated list, and developers of educational instruments use it to create materials of all types. It categorizes hundreds of vocabulary words according to grade placements from primer (pre-grade 1) through grade 13 (post-high school). This was exactly the span desired for the STAR Reading test. No other available vocabulary list provided both the soundness of development and the graded levels across the entire elementary through high school range.

Item Development

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.x test, the item tryout and calibration included all 838 vocabulary items from the STAR Reading 1.x test, plus 836 new vocabulary items created for the STAR Reading 2.x test. It was necessary to write and test about 100 new questions at each grade level 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 2.1.
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.

Table 2.1
Maximum Sentence Length Per Item Grade Level

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

Authentic text passage item specifications

STAR Reading 2.x and higher authentic text passage items are passages of extended text at grade levels 3 through 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.

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

After the grade-level difficulty of a passage was estimated, the passage was searched for occurrences of 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:

- Content material of the passage
- 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.

The STAR Reading 2.x item tryout and calibration included 459 authentic text passage items. About 40 questions at each grade level from 3 through 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.x 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 multi-sentence 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. They do not, however, meet both conditions.
3. The paragraph provides sufficient context clues for students to determine the appropriate answer choice. Average sentence length within the paragraphs is eight to 16 words depending on the item's grade level. Total passage length ranges from 27 to 107 words, based on the average reading speed of each grade level, as shown in Table 2.2.

4. Answer choices for authentic text passage items are EDL Core Vocabulary 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 vocabulary level or one grade below.

Table 2.2
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

Item and Scale Calibration

STAR Reading 3.x RP and higher uses the same bank of calibrated items as STAR Reading 2.x. This chapter summarizes the psychometric research and development undertaken to prepare a large pool of calibrated reading test questions for use in the STAR Reading 2.x test, and to link STAR Reading 2.x scores to the original STAR Reading 1.x score scale. This research took place in two stages: item calibration and score scale calibration. These are described in their respective sections below.

Item Calibration

The previous chapter described the design and development of the STAR Reading 2.x test items. 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.x development, a large-scale item calibration program was conducted in the spring of 1998. The STAR Reading 2.x 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.x 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 first through twelfth grade.

Sample description

The data collection phase of the STAR Reading 2.x 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 3.1 provides the numbers of students in each grade who participated in the study.

Table 3.2 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 3.2 illustrates, the tryout sample approximated the national school population fairly well.

Table 3.1
Numbers of Students Tested by Grade
STAR Reading 2.x Item Calibration Study — Spring 1998

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

Table 3.2
Sample Characteristics
STAR Reading 2.0 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%
	Nonpublic	10%	14%
School Type & District Enrollment	Public		
	<200	17%	15%
	200-499	19%	21%
	500-2,000	27%	25%
	>2,000	28%	24%
	Nonpublic	10%	14%

Table 3.3 provides information about the ethnic composition of the calibration sample. As Table 3.3 shows, the students participating in the calibration sample closely approximate the national school population.

Table 3.3
Ethnic Group Participation
STAR Reading 2.0 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.x 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 3.4. 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 3.4 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 3.4
Calibration Test Forms Design by Test Level
STAR Reading 2.x 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.x 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 3.4 (varying between 11 and 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 the majority of questions in the STAR Reading 2.x calibration study already had some performance data on them. All of the questions from the STAR Reading 1.x item bank were included, as were many items that were previously field tested, but were not included in the STAR Reading 1.x test.

Following extensive quality control checks, the STAR Reading 2.x 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%) indicate that the item is difficult since only a small percentage of students answered it correctly. High p-values (such as 90%) 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.

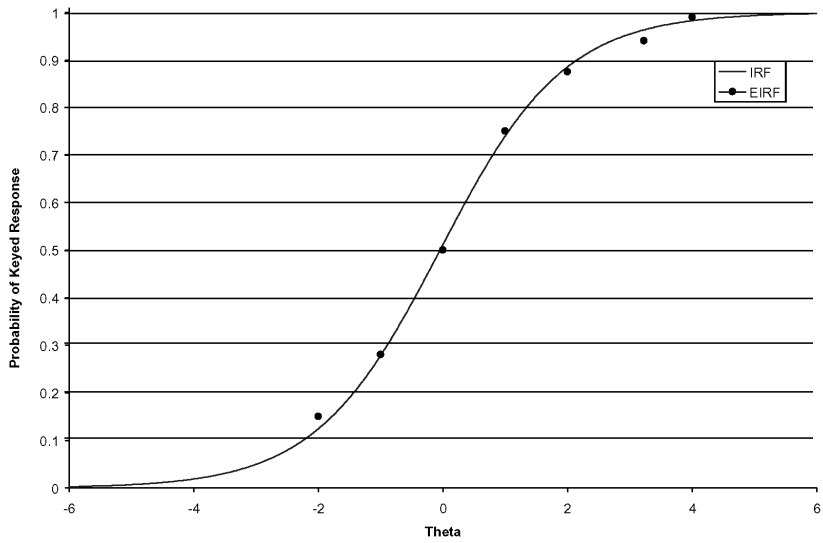
Item response function

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

Within IRT, the probability of answering an item correctly is a function of the student's ability and the difficulty of the item. Since IRT places the item difficulty and student ability on the same scale, this relationship can be represented graphically in the form of an Item Response Function (IRF). Plotting the IRF (as shown by the solid line in Figure 1), one sees that the result is an S-shaped (ogive) function. The difficulty of the item constitutes the horizontal axis; the vertical axis is the probability of a correct response. For any specific item, the probability of answering the item correctly for students whose ability level is much less than the item's difficulty level is low. As the student's ability level increases, relative to the item's

difficulty level, the probability of answering the item correctly increases until the probability nears 1.0. The midpoint, or point of inflection, of the IRF is the difficulty level of the item and is the point where a student with exactly the same level of ability as the item is difficult would be expected to have a 50% chance of answering the item correctly. It is at or near this level that measurement of student achievement is optimal from the perspective of information theory.

Figure 1. Example of Item Statistics Database Presentation of Information



Calibration of test items by IRT methods estimates the IRT difficulty parameter for each test item and places all of the items 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. For purposes of the STAR Reading 2.x calibration research, two different "fit" measures (both unweighted and weighted) were computed. Empirical Item Response Functions (EIRF) for the data were also determined. The EIRF is obtained by grouping students who received that item into groups with similar ability levels and then plotting the proportion of students in each group who answered the item correctly (represented by the round dots in Figure 1) over the mean ability level for that group. 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 study:

- 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.x 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.x item bank.

Item selection was done based on the following criteria. Items were eliminated when:

- Item-total correlation (item discrimination) was < .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

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 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 through 12.

Table 3.5 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 3.5
Calibration Test Item Summary Information by Test Level
STAR Reading 2.x 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	.56	-3.61	-2.36
B	2	274	3,911	78	88	.55	-2.35	-0.07
C	3	274	3,468	76	89	.51	-1.60	0.76
D	4	274	3,340	69	81	.51	-0.14	1.53
E	5-6	343	4,046	62	73	.47	1.02	2.14
F	7-9	343	3,875	68	76	.48	2.65	4.00
G	10-12	366	4,941	60	60	.37	4.19	4.72

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 2.x 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 2.x spring 1998 national item calibration study were 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 2.x national item calibration study.

Like STAR Reading 2.x, the STAR Reading 3.x RP and higher test is a fixed-length, 25-item, computer-adaptive test. Students who have not taken a STAR Reading 2.x, 3.x, or 4.x RP 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 2.x, 3.x, or 4.x RP 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 3.x RP and higher 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 grade 1 and 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 six-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 3.x RP and Higher Test

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.x 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 standard error of measurement (SEM). Unlike a conventional paper-and-pencil test, the SEM values for the STAR Reading test are unique for each student. SEM 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.x RP and higher test (grades 1–12). 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 2.x norming study described in "Norming" on page 26.

Scale Calibration

The outcome of the item calibration study described above was a sizeable bank of test items suitable for use in the STAR Reading 2.x test, with an IRT difficulty scale parameter for each item. The difficulty scale itself was devised such that it spanned a range of item difficulty from Kindergarten through grade 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 2.x norming, described in "Norming" on page 26, 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 1.x software. STAR Reading 1.x Scaled Scores ranged from 50 to 1350, in integer units. The relationship of those Scaled Scores to the STAR Reading 2.x IRT ability scale was expected to be direct, but not necessarily linear. For continuity between STAR Reading 1.x and STAR Reading 2.x scoring, it was desirable to be able to report STAR Reading 2.x scores on the same scale used in STAR Reading 1.x scores. To make that possible, a scale linking study was undertaken in conjunction with STAR Reading 2.x norming. At every grade from 1 through 12, a portion of the norming sample was asked to take both versions of the STAR Reading test: 1.x and 2.x. 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.x IRT ability scores into equivalent STAR Reading 1.x Scaled Scores.

The linking study

4,589 students from around the country, spanning all 12 grades, participated in the linking study. Linking study participants took both STAR Reading 1.x and STAR Reading 2.x 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.

Table 3.6
Summary Statistics of STAR Reading 1.x and 2.x Scores from the Linking Study, by Grade – Spring 1999 (N = 4,551 students)

Grade Level	Sample Size	STAR Reading 1.x Scaled Scores		STAR Reading 2.x IRT Ability Scores		STAR Reading 2.x Equiv. 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
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

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.x Scaled Score equivalents range from 0 to 1400; the same scale is used for STAR Reading 3.x RP and higher.[1]

Summary statistics of the test scores of the 4,551 cases included in the linking analysis are listed in Table 3.6. The table lists actual STAR Reading 1.x Scaled Score means and standard deviations, as well as the same statistics for STAR Reading 2.x IRT ability estimates and equivalent Scaled Scores calculated using the conversion table from the linking study. Comparing the STAR Reading 1.x Scaled Score means to the IRT ability score means illustrates how different the two metrics are. Comparing the STAR Reading 1.x Scaled Score means to the STAR Reading 2.x Equivalent Scaled Scores in the rightmost two columns of Table 3.6 illustrates how successful the scale linking was.

1. Data from the linking study made it clear that STAR Reading 2.x software measures ability levels extending beyond the minimum and maximum STAR Reading 1.x Scaled Scores. In order to retain the superior bandwidth of STAR Reading 2.x software, extrapolation procedures were used to extend the Scaled Score range below 50 and above 1350.

Table 3.7 contains an excerpt from the IRT ability to Scaled Score conversion table that was developed in the course of the linking study.

Table 3.7
Example IRT Ability to Equivalent Scaled Score Conversions

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
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

Norming

STAR Reading 3.x and higher uses the same norms as STAR Reading 2.0. This chapter describes the norming of STAR Reading 2.0.

Sample Characteristics

The norming of the STAR Reading 2.0 computer-adaptive test occurred in Spring 1999. To obtain a sample representative of the U.S. school population, the selection of participating schools focused on stratifying the U.S. school population based on three key variables. These variables, in increasing order of importance, included the following:

- *Geographic Region.* Using the categories established by the National Education Association, schools fell into four regions: Northeast, Midwest, Southeast, and West.
- *School System and Per-Grade District Enrollment.* Statistics distributed by Quality Education Data, Inc. (1998) identified public and nonpublic schools. Public schools were categorized into four groups based on their per-grade district enrollment: fewer than 200 students, 200-499 students, 500-1,999 students, and 2,000 or more students.
- *Socioeconomic Status.* Using the Orshansky Index from Quality Education Data (1998), public schools were categorized based on the proportion of students in the district who fall below the federal poverty level. As a result, schools were identified as being either of High, Average, or Low socioeconomic status. (Since socioeconomic data were not available for nonpublic schools, this classification did not include them.)

Although other collected data describe the norming sample more fully, these factors were the primary basis for establishing an appropriate sampling frame. The sampling frame formed a 52-cell matrix (four regional zones x four public-school enrollment groups x three socioeconomic categories, plus four regional cells for nonpublic schools). All schools in the U.S. were categorized into one of the 52 cells, and participation was requested from sufficient numbers of schools to complete the desired sample.

In March 1999, the schools that agreed to participate in the norming program received a version of STAR Reading 2.x software designed to gather norming data. This version of the program captured the test scores for each of the students participating in the norming program.

These schools were asked to test their students within a three-week window during April 1999. In some cases, this time frame was extended slightly to make it possible for more schools to participate.

The final norming sample included a nationally representative mix of approximately 30,000 students from 269 schools. (Appendix A lists the name, location, and region of every school that participated in this and other phases of development.) These schools represented 47 states across the United States. Table 4.1 summarizes the sample according to each of the variables used to select and refine the norming group.

Table 4.1
Sample Characteristics
STAR Reading Norming Program — Spring 1999 (N=29,627 students)

		Students	
		National %	Sample %
Geographic Region	Northeast	20.3%	14.3%
	Midwest	23.8%	29.6%
	Southeast	24.0%	23.3%
	West	31.9%	32.8%
District Socioeconomic Status	Low: 33-99%	30.0%	30.8%
	Average: 16-32%	29.2%	28.5%
	High: 1-15%	31.2%	31.2%
School Type & District Enrollment	Public		
	<200	16.9%	22.2%
	200-499	19.0%	23.0%
	500-1999	26.7%	23.7%
	>1999	27.8%	21.5%
	Nonpublic	9.6%	9.5%

In addition to the main sampling variables summarized in Table 4.1, other information about the sample schools was collected. Although it was not used to select or adjust the norming sample, this information is summarized in Tables 4.2 through 4.4 to further describe the schools and students that make up the norms. These tables also include national figures based on 1998 data provided by Quality Education Data, Inc.

Table 4.2
School Locations
STAR Reading Norming Program — Spring 1999 (N=269 schools, 29,627 students)

	Schools		Students	
	National %	Sample %	National %	Sample %
Urban	27.9%	30.7%	33.2%	31.6%
Suburban	47.3%	42.4%	51.0%	46.4%
Rural	21.1%	25.4%	13.6%	20.5%
Unclassified	3.7%	1.5%	2.3%	1.6%

Table 4.3
Nonpublic School Affiliations
STAR Reading Norming Program — Spring 1999 (N=31 schools, 2817 students)

	Schools		Students	
	National %	Sample %	National %	Sample %
Catholic	38.2%	48.4%	53.9%	52.7%
Other	61.8%	51.6%	46.1%	47.3%

The STAR Reading norming version also gathered gender and ethnic data on the participating students. Various subgroup analyses were conducted to compare the performance among groups of students. These analyses were done using first-time test-takers participating in the norming program. As is typical of reading achievement tests, females slightly outperformed males on the STAR Reading test, especially in grades before high school. And, as is typical for essentially all measures of educational achievement, from standardized tests to teacher-assigned grades, minority students typically scored lower than their white grade peers, with differences in performance generally on the order of one-half to one standard deviation. Given all of these data, it is clear that the STAR Reading norming sample very closely approximates the distribution of the entire U.S. school population in grades 1-12. Thus, STAR Reading norms provide meaningful national comparison data.

Table 4.4
Ethnic Group and Gender Participation
STAR Reading Norming Program — Spring 1999 (N=29,627 students)

		Students	
		National %	Sample %
Ethnic Group	Asian	3.4%	1.3%
	Black	15.1%	9.0%
	Hispanic	12.2%	8.2%
	Native American	0.9%	0.8%
	White	59.0%	41.2%
	Unclassified	9.3%	39.5%
Gender	Female	Not available	49.4%
	Male	Not available	50.6%
	Response Rate		69.5%
Title I Participation		28.0%	27.1%

Note that the Sample percentage for Title I is based on the school Title I percentage reported values weighted for sample size. Title I percentage is only reported for public schools.

Test Administration

All participants in STAR Reading 2.x norming took the test in computer-administered, adaptive form during the spring of 1999. Some students in the normative sample also participated in one of two collateral studies conducted at close points in time: a test-retest reliability study and the linking study described in "Item and Scale Calibration" on page 14. For the test-retest study, students took a second STAR Reading 2.x test within a few days of the first. Conditions for administering the second test were identical to the first, except that items administered to a given student during the first test were not used again. For the linking study, students took two different versions of the STAR Reading test, in counterbalanced order within a few days of one another. On one occasion, the STAR Reading 1.x test was administered; on the other occasion, the STAR Reading 2.x norming version was used.

Data Analysis

After the participating schools tested their students, they returned their student test data on floppy disks for analysis. For schools that participated in the test-retest reliability study, only the scores from the test administered first were used for the development of the norm-referenced scores; scores on the second test were used only in the analyses of test-retest reliability. For schools that participated in the linking study of the STAR Reading 1.x and STAR Reading 2.x tests, all of the STAR Reading 2.x scores were included in the norms development, regardless of which test form was taken first.

The normative data (Percentile Ranks and Grade Equivalents) were based on the frequency distributions of the IRT ability estimates (thetas) of the students' first tests; tables of norms presented in this manual and used in the STAR Reading 2.x software contain the Scaled Score equivalents to those IRT ability estimates.

In order to ensure that the norming sample was nationally representative at each grade level, the norming data were statistically weighted to maximize the correspondence with the U.S. school population. The weighting factors were based on the number of students in each sampling cell and the national proportion of the U.S. school population that constituted each sampling cell:

- For grades 1 through 8, the student test results were weighted first by geographic region, and then by the type, size, and the socioeconomic status of the school system.
- The test results for grades 9 through 12, where sample sizes were smaller, were weighted first by geographic region and then by the socioeconomic status of the school system.

Table 4.5 presents the weighted and unweighted Scaled Score data by grade for 29,169[2] students included in the norms analysis. As the tabled values indicate, the weighted and unweighted sample sizes did not differ from each other, but the weighting did make some difference in the means, standard deviations, and median scores at some grades. Consequently, the STAR Reading 2.x norms are based on weighted score frequencies. Comparison of these norms with the STAR Reading 1.x norms indicated that the central tendency of the new and old norms was similar at all grades except 1 and 2. At those two grades, the STAR Reading 2.x median scores were approximately two tenths of a standard deviation higher than the STAR

2. There were 29,627 cases in the STAR Reading 2.0 norms sample; 458 with outlier scores and corrupted data were not included in the norms calculations.

Reading 1.x scores, an increase too large to be credible. It was highly unlikely that children's reading skills had improved that dramatically in the three years between the STAR Reading 1.x norming and the STAR Reading 2.x norming. After careful deliberation, the decision was made to retain the old norms for grades 1 and 2. Consequently, while tables in this chapter reflect the STAR Reading 2.x sample data, the STAR Reading 2.x software, and tables in subsequent chapters, incorporate STAR Reading 1.x norms for grades 1 and 2.

Table 4.5
Comparison of Unweighted (U) and Weighted (W) Scaled Scores
STAR Reading Norming Program — Spring 1999 (N=29,169 students)

Grade Level	Sample Size		Scaled Score Means		Scaled Score Standard Deviations		Scaled Score Medians	
	U	W	U	W	U	W	U	W
1	2,703	2,703	205	204	124	126	177	174
2	3,292	3,292	344	349	142	142	341	346
3	2,923	2,923	437	432	164	167	443	435
4	3,720	3,720	522	519	197	198	510	506
5	3,177	3,177	645	633	231	232	617	601
6	2,793	2,793	757	768	269	266	734	749
7	3,395	3,395	845	846	274	273	842	843
8	2,838	2,838	917	935	285	283	921	944
9	1,855	1,856	962	969	286	285	988	1,045
10	1,124	1,122	1,027	1,021	285	289	1,108	1,104
11	755	755	1,089	1,075	267	272	1,203	1,147
12	594	594	1,134	1,089	261	285	1,249	1,183

These norming procedures resulted in empirical, nationally representative scores for the STAR Reading 2.x computer-adaptive test. These norm-referenced scores correspond directly with the time period during which the norming study was conducted (the month of April). Norm-referenced scores for each month of the school year were then determined through a process of interpolation between the adjacent empirical norms for each Scaled Score point, assuming equal growth between adjacent points in time. This allows the STAR Reading 2.x test to provide normative information that is most relevant, regardless of the specific time period in which schools administer the test to their students.

Grade Equivalent (GE) scores within the normative sample were defined as the median (50th percentile) Scaled Scores at each grade; as the mean test date was in the month of April, these empirical median scores constitute the GE scores for month seven of each grade. GE scores for other time periods were determined by interpolation.

Scaled Score to Percentile Rank conversion tables for the empirical norming period are presented in Table 7.2 on page 59. The Scaled Score to Grade Equivalent conversion table is presented in Table 7.1 on page 54.

Score Definitions

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 through June and assigns increment values of .0 through .9 to these months. Table 5.1 on page 33 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, make sure the following school year is set as the current school year in the Renaissance Place program at the time you 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, make sure the school year that has just ended is set as the current school year in the Renaissance Place program at the time you 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 your school follows the standard school calendar used in STAR Reading software and you will not be testing in the summer, assigning the appropriate grade placements for your students is relatively easy. However, if you are going to test your students in July or August — whether it is for a summer reading

program or because your 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, you must determine when to set your next school year as your current school year, and thereby advance students from one grade to the next. In most cases, you can use the guidelines above.

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 you discover that a student has tested with an incorrect grade placement assignment (use the Growth, Snapshot, Summary, or Test Record Report to find out the grade placement), the procedures outlined on page 36 in the discussion about Table 7.2 can be used to arrive at corrected estimates for the student's Percentile Rank and Normal Curve Equivalent scores.

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 his 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 software is the Instructional Reading Level, which compares a student's test performance to updated vocabulary lists that were based on the EDL Core Vocabulary. The Instructional Reading Level is the highest grade level at which the student is estimated to know at least 80 percent of the vocabulary words.
- *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 1999 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.x 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 14. STAR Reading 2.x and higher Scaled Scores range from 0 to 1400.

Grade Equivalent (GE)

A Grade Equivalent (GE) indicates the normal grade placement of students for whom a particular score is typical. For example, the median (typical) Scaled Score obtained by third-graders in the seventh month (April) during STAR Reading 2.x norming was 432. Thus, the Grade Equivalent score for anyone receiving a Scaled Score of 432 is 3.7.

STAR Reading Grade Equivalents range from 0.0 to 12.9+. Because the GE scale expresses individual "months" in tenths, the scale does not cover the summer months. Table 5.1 indicates how the decimalized GE tenths correspond to the various calendar months. Since the STAR Reading 2.x norming took place during the seventh month of the school year (April), GEs ending in .7 are empirically based; in other words, they provide conversions based on actual normative medians. All other portions of the scale are formed by fitting a curve to the grade-by-grade medians and finding Scaled Scores that fit the curve. Table 7.1 on page 54 contains the Scaled Score to GE conversions.

Table 5.1
Incremental Grade Placement Values Per Month

Month	Decimal Increment
July	0.00 or 0.99 (depends on the current school year set in Renaissance Place)
August	0.0 or 0.99 (depends on the current school year set in Renaissance Place)
September	0.0
October	0.1
November	0.2
December	0.3
January	0.4
February	0.5
March	0.6
April	0.7
May	0.8
June	0.9

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

Comparing the STAR Reading Test with 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 performance can be appropriately compared to that of a typical seventh-grader in the sixth month of the school year.

Instructional Reading Level (IRL)

The Instructional Reading Level is a criterion-referenced score that is an estimate of the most appropriate level of reading material for instruction. In other words, IRLs tell you the reading level at which students can recognize words and understand 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), and PHS (Post-High School).

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 answer 80% or more of the items correctly. For example, if a student is able to answer 80% of the seventh-grade test items correctly, but only 60% of the eighth-grade items correctly, he or she would have a seventh-grade Instructional Reading Level. STAR Reading 2.x and higher software 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 answer correctly at each grade level.

Special IRL scores

If a student's STAR Reading 2.x or 3.x RP and higher performance 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 nonreader. 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% 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% 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% 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 to 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 1119 corresponds to a PR of 80, and a Scaled Score of 1222 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 7.2 on page 59 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 through 12.9. Because STAR Reading norming occurred in the seventh month of the school year (April), the seventh-month values for each grade are empirically based. The remaining monthly values were estimated by interpolating between the empirical points. The table also includes a column representing students who are just about to graduate from high school.

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 *Grade Placement* on page 31 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, enter Table 7.2 in the 3.0 grade placement column and proceed down the table until you find the student's Scaled Score or the next-higher value in the table. Then, 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, 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 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 7.2 (page 59), converting them to NCE values using Table 7.3 (page 63), interpolating between the NCE values, and then converting the interpolated NCE value back to a PR value using Table 7.4 (page 64).

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 to 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 interpolating test scores, and 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 7.3 on page 63 provides the NCEs corresponding to integer PR values and facilitates the conversion of PRs to NCEs. Table 7.4 on page 64 provides the conversions from NCE to PR. The NCE values are given as a range of scores that convert to the corresponding PR value.

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 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 7.5 on page 66 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 all books included in the system. 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 *Research Foundation for Reading Renaissance Goal Setting* (2003), which is published by Renaissance Learning.

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 5.2 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 below the 25th percentile.

Table 5.2
Diagnostic Code Values by Percentile Rank

Grade	Diagnostic Code	
	PR > 25	PR <=25
0.0 to 0.7	01A	01B
0.8 to 1.7	02A	02B
1.8 to 2.7	03A	03B
2.8 to 3.7	04A	04B
3.8 to 4.7	05A	05B
4.8 to 5.7	06A	06B
5.8 to 6.7	07A	07B
6.8 to 8.7	08A	08B
8.8 to 13.0	09A	09B

A Ph.D. reading specialist developed the diagnostic codes and accompanying text using standard scope and sequence paradigms from the field of reading education. Two other Ph.D.s in reading education (a tenured professor and a Title I supervisor for a large metropolitan school district) also reviewed the codes and text descriptions. These reviewers found:

1. The diagnostic information succinctly characterizes readers at each stage of development and across grade levels K-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.

Reliability and Validity

Reliability is the extent to which a test yields consistent results from one administration to another and from one test form to another. Tests must yield consistent results in order to be useful. Because the STAR Reading 2.x test is a computer-adaptive test, many of the typical methods used to assess reliability using internal consistency methods (such as KR-20 and coefficient alpha) are not appropriate.

There are, however, four direct methods that can be used to estimate the reliability of the STAR Reading computer-adaptive test: the split-half method, the test-retest method, the alternate forms method, and the estimation of generic reliability. In the course of the STAR Reading 2.x norming, data were collected to allow all four of these methods to be applied. The results apply to STAR Reading 2.x and higher.

Split-half Reliability Analysis

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

In internal simulation studies, the split-half method provided accurate estimates of the internal consistency reliability of adaptive tests, and so it has been used to provide estimates of STAR Reading 2.x and higher reliability. These split-half reliability coefficients are independent of the generic reliability approach discussed below and more firmly grounded in the item response data. Split-half scores were based on the first 24 items of the STAR Reading 2.0 norming test; scores based on the odd- and the even-numbered items were calculated. The correlations between the two sets of scores were corrected to a length of 25 items, yielding the split-half reliability estimates displayed in the fourth column of Table 6.1 on page 42.

Test-Retest Reliability Study

The test-retest study provided estimates of STAR Reading 2.x 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 2.x test-retest 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 use of different items for tests one and two makes the test-retest study a kind of alternate forms reliability study, but that term is reserved for another study, described below.) 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 test-retest study's estimates indicate.

The test-retest reliability estimates for the STAR Reading 2.x 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. The final sample for the STAR Reading 2.x test-retest reliability study consisted of a total of 2,095 students — a reasonable number of students for these kinds of analyses.

It is important to note that very little time elapsed between the first and second administrations of the students' tests. The median date of administration for the first test (across grades) was April 20, 1999, while the median date for administration of the second test was April 27, 1999. Consequently, it is safe to assume that no measurable growth in reading ability or achievement occurred between the two testing occasions. Unlike the operational form of STAR Reading 2.x and higher software, in which the starting ability estimate for subsequent testing sessions is dependent upon previous test scores, the test-retest reliability version of STAR Reading 2.x software was constrained to start both tests at the same point. This helped maximize the parallelism of the two tests.

Reliability coefficients estimated from the test-retest study are provided in the fifth column of Table 6.1 on page 42. The test-retest coefficients listed there are corrected correlation coefficients. The observed correlations have been corrected for differences between the score variances of this study's sample and the weighted normative sample; the corrections were very small, and worked in both directions, increasing some reliability estimates and decreasing others.

Correlation coefficients range from -1 to +1, where -1 is a perfect negative correlation and +1 is a perfect positive correlation. As Table 6.1 shows, the test-retest reliability estimated over all 12 grades was .94. Estimates by grade, which are smaller because score variances within grades are smaller, range from .79 to .91. Their average is .85 — quite high for reliability estimates of this type. 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 2.x 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.

Alternate Forms Linking Study

The linking study described in "Item and Scale Calibration" provided an opportunity to develop estimates of STAR Reading alternate forms reliability. Students in this study took both a STAR Reading 2.x test and an original STAR Reading 1.x test, with an interval of days between tests. Order of administration was counterbalanced, with some students taking the STAR Reading 1.x test first, and the others taking the new STAR Reading 2.x test first. The correlations between scores on the two tests were taken as estimates of alternate forms reliability. These correlation coefficients should be similar in magnitude to those of the test-retest study, but perhaps somewhat lower because the differences between versions 1.x and 2.x, summarized in the Introduction, contribute additional sources of measurement error variance. These differences are material: STAR Reading 1.x tests are longer than the 25-item STAR Reading 2.x tests, are variable-length rather than fixed-length, are more homogeneous in content (consisting solely of vocabulary-in-context items), and are not based on the IRT technology that is the psychometric foundation of the STAR Reading 2.x test.

The alternate forms reliability estimates from the linking study are shown for grades 1 through 12 in the rightmost column of Table 6.1 on page 42. As the data in this table indicate, the correlation was .95 for the overall sample of 4,551 students. By grade, sample sizes ranged from 63 to 772, with most samples larger than 200 cases. The reliability coefficients within grade ranged from .82 to .89. Like the test-retest reliability coefficients, their average is .85. The magnitude of these correlations speaks not only to the reliability of the STAR Reading 2.x and higher tests (and 1.x as well) but also to their equivalence as measures of reading performance.

Generic Reliability Study

The data of the norming study as a whole provided the opportunity to estimate what we refer to here as generic reliability coefficients for the STAR Reading 2.x and higher tests. Estimates of generic reliability are derived from an IRT-based feature of the STAR Reading 2.x and higher tests: individual estimates of measurement error, called conditional SEMs, that are computed along with each student's IRT ability estimate, theta. Item Response Theory, and hence STAR Reading 2.x and higher software, acknowledges that measurement precision and measurement error are not constant, but vary with test score levels. It is possible to estimate the classical reliability coefficient using the conditional SEMs and the variance of the IRT-based observed scores.

Since the classical concept of reliability can be defined as

1-(error variance / total score variance)

we can compute a reliability estimate by substituting the average of the individual student error variances (as the error variance term) and the variance of the students' ability estimates (as the best estimate of total score variance). Like all measures of reliability, this method looks at the proportion of overall score variance that is exclusive of measurement error.

Using this technique with the STAR Reading 2.x norming data resulted in the generic reliability estimates shown in the third column of Table 6.1. Because this method is not susceptible to problems associated with repeated testing and alternate forms, the resulting estimates of reliability are generally higher than the more conservative test-retest and alternate forms reliability coefficients. Estimation of generic reliability also makes use of all the data in the norming study (N = 29,169), not just the subset of the overall sample that participated in the two reliability studies (N = 2,095 and N=4,551). These generic reliability coefficients are, therefore, a more plausible estimate of the actual reliability of the STAR Reading 2.x and higher adaptive tests than are the two more conservative coefficients.

The generic reliability estimates listed in Table 6.1 range from .89 to .92, and vary little from grade to grade. These reliability estimates are 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 2.x test.

Table 6.1
Reliability Estimates from the STAR Reading Norming and Reliability Studies — Spring 1999

Grade	Norming Sample			Test-Retest Sample		Alternate Forms Sample	
	N	Generic Reliability	Split-half Reliability	N	Retest Reliability	N	Alternate Forms Reliability
All				2,095	0.94	4,551	0.95
1	2,703	0.92	0.89	301	0.91	284	0.88
2	3,292	0.90	0.89	287	0.86	772	0.89
3	2,923	0.91	0.89	223	0.87	476	0.86
4	3,720	0.90	0.90	341	0.85	554	0.87
5	3,177	0.90	0.89	264	0.84	520	0.83
6	2,793	0.89	0.90	175	0.82	219	0.82
7	3,395	0.89	0.89	145	0.86	702	0.82
8	2,838	0.89	0.89	125	0.82	545	0.83
9	1,855	0.89	0.91	97	0.90	179	0.87
10	1,124	0.90	0.89	80	0.86	81	0.88
11	755	0.89	0.91	26	0.79	156	0.82
12	594	0.90	0.93	31	0.85	63	0.82

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. For example, a 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% of the time, and within 72 points (twice the SEM) roughly 95% 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 2.x 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/her performance, unlike most printed tests that report the same SEM value for every examinee. Each administration of the test yields a unique SEM that reflects the amount of information estimated to be in the specific combination of items that a student received in his or her individual test. Second, because the STAR Reading test is adaptive, the SEM will tend to be lower than that of a conventional test, particularly at the highest and lowest score levels, where conventional tests' measurement

precision is weakest. Because the adaptive testing process attempts to provide equally precise measurement, regardless of the student's ability level, the average SEMs for the IRT ability estimates are very similar for all students. However, because the transformation of the IRT ability estimates into equivalent Scaled Scores is not linear, the SEMs in the Scaled Score metric are less similar.

Table 6.2 summarizes the average SEM values for the norms sample, overall and by grade level. The third column contains the average IRT score SEMs (multiplied by 100 to eliminate decimals). The fourth column contains average Scaled Score SEMs. As the data indicate, the average IRT score SEMs are nearly constant regardless of grade level. In contrast, the SEMs of the Scaled Scores vary widely by grade, increasing from an average of 37 points at grade 1 to 96 points at grade 7, then decreasing to 76 at grade 12. To illustrate the variability of individual Scaled Score SEMs, the table displays the 5th and 95th percentiles of the SEMs at each grade. The range of SEMs between these two percentile values varies widely, with the largest range — 137 points — at grade 10.

Table 6.2
Standard Errors of Measurement (IRT Scores and Scaled Scores)
STAR Reading Norming Analysis — Spring 1999

Grade	Norming Sample Size	Average IRT Score SEM x100	Average Scaled Score SEM	5th Percentile Scaled Score SEM	95th Percentile Scaled Score SEM
Overall	29,169	49	74	21	137
1	2,703	50	37	5	72
2	3,292	49	49	28	75
3	2,923	48	56	35	100
4	3,720	49	66	37	131
5	3,177	49	80	40	141
6	2,793	50	94	41	148
7	3,395	48	96	43	145
8	2,838	49	95	36	144
9	1,855	48	92	23	143
10	1,124	48	84	3	140
11	755	48	80	3	135
12	594	49	76	3	130

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 — the relevance of the test questions to the attributes supposed to be measured by the test — reading ability, in the case of the STAR Reading test. These content validity issues were discussed in detail in "Content and Item Development" (page 10) and were an integral part of the design and construction of the STAR Reading 2.x and higher test items.

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, the STAR Reading 2.x and higher tests claim to provide an estimate of a child's reading achievement level. Therefore, demonstration of the STAR Reading test's construct validity rests on the evidence that the test in fact provides such an estimate. There are, of course, a number of ways to demonstrate this. Since reading ability varies significantly within and across grade levels and improves as a student's grade placement increases, STAR Reading 2.x and higher scores should demonstrate these anticipated internal relationships; in fact, they do. Additionally, STAR Reading 2.x and higher scores should correlate highly with other accepted procedures and measures that are used to determine reading achievement level; this is external validity.

External Validity

During the STAR Reading 2.x norming study, schools submitted data on how their students performed on several other popular standardized test instruments along with their students' STAR Reading results. This data included more than 12,000 student test results 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* (SAT-9), 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.x scale 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.

Tables 6.3 and 6.4 present the correlation coefficients between the STAR Reading 2.x test and each of the other test instruments for which data were received. Table 6.3 displays "concurrent validity" data, that is, correlations between STAR Reading 2.0 norming study test scores and other tests administered at close to the same time. Tests listed in Table 6.3 were administered during the Spring of 1999, the same quarter in which the STAR Reading 2.0 norming study took place. Table 6.4 displays all other correlations of STAR Reading 2.0 norming tests and external tests; the external test scores were administered at various times prior to Spring 1999, and were obtained from student records.

Each table is presented in two parts, A and B. Part A presents validity coefficients for grades 1 through 6, and part B presents the validity coefficients for grades 7 through 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 varied from .60 to .81; the overall average was .76 for grades 1 through 6, and .68 for grades 7 through 12. The other validity coefficient within-grade averages varied from .60 to .77; the overall average was .73 for grades 1 through 6, and .71 for grades 7 through 12.

The extent that the STAR Reading 2.x test correlates with these tests provides support for STAR Reading construct validity.

While these correlation coefficients are high, they are likely conservative in their estimation of the actual correlation between the STAR Reading test and the other standardized reading tests. The actual relationship between the STAR Reading test and the other tests is likely a bit higher than these estimates indicate. This degree of conservatism results from two factors. First, the standardized test scores reported were from tests administered at points in time that were far different from the administration of the STAR Reading 2.x test; generally, the degree of correlation between two test scores decreases as the time between test administration increases. Second, the collection of the standardized test scores for the validity analyses involved a manual process of teachers transcribing scores for students onto forms printed by the STAR Reading 2.x norming version. Although several safeguards to reduce sources of error were put into place, this procedure was not immune to errors.

The process of establishing the validity of a test is an involved one, and one that usually takes a significant amount of time. Thus, the data collection process and the validation of the STAR Reading test is really an ongoing activity seeking to establish evidence of STAR Reading's validity for a variety of settings. STAR Reading users who collect relevant data are encouraged to contact Renaissance Learning, Inc.

Since correlation coefficients are available for different editions, forms, and dates of administration, many of the tests have several correlation coefficients associated with them. Where test data quality could not be verified, and when sample size was limited, those data were eliminated. Correlations were computed separately on tests according to the unique combination of test edition/form and time when testing occurred. Testing data for other standardized tests administered more than two years prior to spring 1999 were excluded from the analyses since those test results represent very dated information about the current reading ability of students.

In general, these correlation coefficients reflect very well on the validity of the STAR Reading 2.x and higher tests as tools for assessing reading performance. These results, combined with the reliability and SEM estimates, demonstrate quantitatively how well this innovative instrument in reading assessment performs.

Table 6.3-A
Concurrent Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Spring 1999, Grades 1-6. Sample sizes are in the columns labeled "n."

Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
California Achievement Test (CAT)														
/ 5	Spr 99	NCE	93	0.80*	36	0.67*	–	–	34	0.72*	146	0.76*	–	–
Comprehensive Test of Basic Skills (CTBS)														
/ 4	Spr 99	NCE	–	–	–	–	–	–	18	0.81*	–	–	–	–
A-19/20	Spr 99	Scaled	–	–	–	–	–	–	–	–	–	–	8	0.91*
Gates-MacGinitie Reading Test (GMRT)														
2nd Ed., D	Spr 99	NCE	–	–	21	0.89*	–	–	–	–	–	–	–	–
L-3rd	Spr 99	NCE	–	–	127	0.80*	–	–	–	–	–	–	–	–
Iowa Test of Basic Skills (ITBS)														
Form K	Spr 99	NCE	40	0.75*	36	0.84*	26	0.82*	28	0.89*	79	0.74*	–	–
Form L	Spr 99	NCE	–	–	–	–	18	0.7*	29	0.83*	41	0.78*	38	0.82*
Form M	Spr 99	NCE	–	–	–	–	158	0.81*	–	–	125	0.84*	–	–
Form K	Spr 99	Scaled	–	–	58	0.74*	–	–	54	0.79*	–	–	–	–
Form L	Spr 99	Scaled	–	–	–	–	45	0.73*	–	–	–	–	50	0.82*
Metropolitan Achievement Test (MAT)														
7th Ed.	Spr 99	NCE	–	–	–	–	–	–	46	0.79*	–	–	–	–
6th Ed	Spr 99	Raw	–	–	–	–	8	0.58*	–	–	8	0.85*	–	–
7th Ed.	Spr 99	Scaled	–	–	–	–	25	0.73*	17	0.76*	21	0.76*	23	0.58*
Missouri Mastery Achievement Test (MMAT)														
	Spr 99	NCE	–	–	–	–	–	–	–	–	26	0.62*	–	–
North Carolina End of Grade Test (NCEOG)														
	Spr 99	Scaled	–	–	–	–	–	–	–	–	85	0.79*	–	–
Stanford Achievement Test (Stanford)														
9th Ed.	Spr 99	NCE	68	0.79*	–	–	26	0.44*	–	–	–	–	86	0.65*
9th Ed.	Spr 99	Scaled	11	0.89*	18	0.89*	67	0.79*	66	0.79*	72	0.80*	64	0.72*
Terra/Nova														
	Spr 99	Scaled	–	–	61	0.72*	117	0.78*	–	–	–	–	–	–
Texas Assessment of Academic Skills (TAAS)														
	Spr 99	NCE	–	–	–	–	–	–	–	–	–	–	229	0.66*
Woodcock Reading Mastery (WRM)														
	Spr 99		–	–	–	–	–	–	–	–	7	0.68*	7	0.66*

Table 6.3-A (Continued)
Concurrent Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Spring 1999, Grades 1-6. Sample sizes are in the columns labeled "n."
Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

Test Form	Date	Score	1		2		3		4		5		6	
			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	2,466		212		357		490		292		610		505	
Number of coefficients	46		4		7		9		8		10		8	
Average validity	–		0.81		0.79		0.71		0.8		0.76		0.73	
Overall average	0.76													

Table 6.3-B
Concurrent Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Spring 1999, Grades 7-12. Sample sizes are in the columns labeled "n."
Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

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)														
/ 5	Spr 99	NCE	–	–	–	–	59	0.65*	–	–	–	–	–	–
/ 5	Spr 99	Scaled	124	0.74*	131	0.76*	–	–	–	–	–	–	–	–
Iowa Test of Basic Skills (ITBS)														
Form K	Spr 99	NCE	–	–	–	–	67	0.78*	–	–	–	–	–	–
Form L	Spr 99	Scaled	47	0.56*	–	–	65	0.64*	–	–	–	–	–	–
Missouri Mastery Achievement Test (MMAT)														
	Spr 99	NCE	–	–	29	0.78*	19	0.71*	–	–	–	–	–	–
Northwest Evaluation Association Levels Test (NWEA)														
Achieve	Spr 99	NCE	–	–	124	0.66*	–	–	–	–	–	–	–	–
Stanford Achievement Test (Stanford)														
9th Ed.	Spr 99	NCE	50	0.65*	50	0.51*	–	–	–	–	–	–	–	–
9th Ed.	Spr 99	Scaled	70	0.70*	68	0.80*	–	–	–	–	–	–	–	–
Test of Achievement and Proficiency (TAP)														
	Spr 99	NCE	–	–	–	–	6	0.42	13	0.80*	7	0.60	–	–
Texas Assessment of Academic Skills (TAAS)														
	Spr 99	NCE	–	–	–	–	–	–	43	0.60*	–	–	–	–

Table 6.3-B (Continued)
Concurrent Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Spring 1999, Grades 7-12. Sample sizes are in the columns labeled "n."
Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
Wide Range Achievement Test 3 (WRAT3)														
	Spr 99		–	–	17	0.81*	–	–	–	–	–	–	–	–
Summary														
Grade(s)		**All**	**7**		**8**		**9**		**10**		**11**		**12**	
Number of students		989	291		419		216		56		7		0	
Number of coefficients		18	4		6		5		2		1		0	
Average validity		–	0.66		0.72		0.64		0.7		0.6		–	
Overall average		0.68												

Table 6.4-A
Other External Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1-6. Sample sizes are in the columns labeled "n."
Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

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*

Table 6.4-A (Continued)
Other External Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1-6. Sample sizes are in the columns labeled "n."
Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
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*	–	–	–	–
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*	–	–	–	–	–	–	–	–

Table 6.4-A (Continued)
Other External Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1-6. Sample sizes are in the columns labeled "n."
Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
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*	–	–	–	–	–	–	–	–
Terra/Nova														
	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												

Table 6.4-B
Other External Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 7-12. Sample sizes are in the columns labeled "n."
Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

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*	–	–	–	–	–	–
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*	–	–

Table 6.4-B (Continued)
Other External Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 7-12. Sample sizes are in the columns labeled "n."
Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level.

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
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*	–	–	–	–
Terra/Nova														
	Fall 97	NCE	103	0.69*	–	–	–	–	–	–	–	–	–	–
	Spr 98	Scaled	–	–	87	0.82*	–	–	21	0.47*	–	–	–	–
	Fall 98	NCE	35	0.69*	32	0.74*	–	–	–	–	–	–	–	–
Test of Achievement and Proficiency (TAP)														
	Spr 97	NCE	–	–	–	–	–	–	–	–	36	0.59*	–	–
	Spr 98	NCE	–	–	–	–	–	–	41	0.66*	–	–	43	0.83*
Texas Assessment of Academic Skills (TAAS)														
	Spr 97	TLI	–	–	–	–	–	–	–	–	–	–	41	0.58*
Wide Range Achievement Test 3 (WRAT3)														
	Spr 98		9	0.35	–	–	–	–	–	–	–	–	–	–
	Fall 98		–	–	–	–	16	0.80*	–	–	–	–	–	–
Wisconsin Reading Comprehension Test														
	Spr 98		–	–	–	–	–	–	63	0.58*	–	–	–	–
Summary														
Grade(s)		**All**	**7**		**8**		**9**		**10**		**11**		**12**	
Number of students		3,158	1,016		529		733		398		317		165	
Number of coefficients		60	18		15		10		8		5		4	
Average validity		–	0.71		0.72		0.75		0.60		0.69		0.77	
Overall average		0.71												

Meta-Analysis of the STAR Reading Validity Data

Meta-analysis is a set of statistical procedures that combines results from different sources or studies. When applied to a set of correlation coefficients that estimate test validity, meta-analysis combines the observed correlations and sample sizes to yield estimates of overall validity, as well as standard errors and confidence intervals, both overall and within grades. To conduct a meta-analysis of the STAR Reading validity data, the 223 correlations displayed in Tables 6.3 and 6.4 were combined and analyzed using a fixed effects model for meta-analysis. The results are displayed in Table 6.5. The table lists results for the correlations within each grade, as well as results with all twelve grades' data combined. For each set of results, the table lists an estimate of the true validity, a standard error, and the lower and upper limits of a 95 percent confidence interval for the validity coefficient.

Using the pilot study data, the overall estimate of the validity of STAR Reading is .72, with a standard error of .005. The true validity is estimated to lie within the range of .71 to .73, with a 95 percent confidence level. The probability of observing the 223 correlations reported in Tables 6.3 and 6.4, if the true validity were zero, is virtually zero. Because the 223 correlations were obtained with widely different tests, and among students from twelve different grades, these results provide support for the validity of STAR Reading as a measure of reading ability.

Table 6.5
Results of the Meta-Analysis of STAR Reading 2.x Correlations with Other Tests

Grade	Effect Size		95% Confidence Level	
	Validity Estimate	Standard Error	Lower Limit	Upper Limit
1	0.77	0.02	0.72	0.81
2	0.72	0.02	0.68	0.74
3	0.75	0.01	0.73	0.78
4	0.75	0.01	0.73	0.77
5	0.75	0.01	0.72	0.77
6	0.71	0.01	0.68	0.74
7	0.70	0.01	0.67	0.72
8	0.72	0.02	0.69	0.75
9	0.72	0.02	0.69	0.75
10	0.61	0.03	0.55	0.67
11	0.70	0.03	0.64	0.75
12	0.74	0.03	0.69	0.79
All	0.72	0.00	0.71	0.73

Conversion Tables

Table 7.1
Scaled Score to Grade Equivalent Conversions

SS Range		Grade Equivalent
Low	High	
0	45	0.0
46	50	0.1
51	55	0.2
56	58	0.3
59	60	0.4
61	63	0.5
64	65	0.6
66	68	0.7
69	71	0.8
72	79	0.9
80	82	1.0
83	86	1.1
87	89	1.2
90	96	1.3
97	105	1.4
106	121	1.5
122	141	1.6
142	159	1.7
160	176	1.8
177	194	1.9
195	212	2.0
213	229	2.1
230	247	2.2
248	266	2.3

Table 7.1 (Continued)
Scaled Score to Grade Equivalent Conversions

SS Range		Grade Equivalent
Low	High	
267	283	2.4
284	302	2.5
303	322	2.6
323	333	2.7
334	343	2.8
344	354	2.9
355	364	3.0
365	372	3.1
373	383	3.2
384	395	3.3
396	407	3.4
408	421	3.5
422	434	3.6
435	442	3.7
443	449	3.8
450	455	3.9
456	461	4.0
462	466	4.1
467	473	4.2
474	481	4.3
482	490	4.4
491	497	4.5
498	505	4.6
506	514	4.7
515	522	4.8
523	531	4.9
532	542	5.0
543	553	5.1

Table 7.1 (Continued)
Scaled Score to Grade Equivalent Conversions

SS Range		Grade Equivalent
Low	High	
554	560	5.2
561	569	5.3
570	579	5.4
580	588	5.5
589	600	5.6
601	612	5.7
613	624	5.8
625	637	5.9
638	650	6.0
651	664	6.1
665	678	6.2
679	693	6.3
694	710	6.4
711	726	6.5
727	748	6.6
749	763	6.7
764	772	6.8
773	780	6.9
781	788	7.0
789	796	7.1
797	805	7.2
806	814	7.3
815	824	7.4
825	833	7.5
834	842	7.6
843	852	7.7
853	864	7.8
865	877	7.9

Table 7.1 (Continued)
Scaled Score to Grade Equivalent Conversions

SS Range		Grade Equivalent
Low	High	
878	888	8.0
889	897	8.1
898	904	8.2
905	910	8.3
911	919	8.4
920	930	8.5
931	943	8.6
944	950	8.7
951	959	8.8
960	966	8.9
967	972	9.0
973	978	9.1
979	987	9.2
988	1,001	9.3
1,002	1,016	9.4
1,017	1,032	9.5
1,033	1,044	9.6
1,045	1,050	9.7
1,051	1,055	9.8
1,056	1,060	9.9
1,061	1,066	10.0
1,067	1,071	10.1
1,072	1,080	10.2
1,081	1,089	10.3
1,090	1,095	10.4
1,096	1,099	10.5
1,100	1,103	10.6
1,104	1,106	10.7

Table 7.1 (Continued)
Scaled Score to Grade Equivalent Conversions

SS Range		Grade Equivalent
Low	High	
1,107	1,110	10.8
1,111	1,115	10.9
1,116	1,120	11.0
1,121	1,124	11.1
1,125	1,129	11.2
1,130	1,133	11.3
1,134	1,137	11.4
1,138	1,142	11.5
1,143	1,146	11.6
1,147	1,151	11.7
1,152	1,155	11.8
1,156	1,160	11.9
1,161	1,163	12.0
1,164	1,166	12.1
1,167	1,170	12.2
1,171	1,173	12.3
1,174	1,176	12.4
1,177	1,179	12.5
1,180	1,182	12.6
1,183	1,185	12.7
1,186	1,189	12.8
1,190	1,192	12.9
1,193	1,400	12.9+

Table 7.2
Scaled Score to Percentile Rank Conversions

	Grade Placement											
PR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	50	59	74	76	99	147	201	256	318	332	342	386
2		62	86	86	109	209	265	310	363	377	378	422
3	62	65	93	109	178	262	316	363	401	413	422	461
4		67	98	145	211	284	357	401	439	445	461	496
5		69	112	166	236	316	382	439	463	471	496	526
6	64	70	118	178	252	319	400	458	477	492	526	555
7		72	124	194	263	349	424	472	495	513	541	580
8		73	136	215	283	360	442	486	513	528	556	600
9	65	75	146	230	294	370	452	503	528	546	583	620
10		77	158	241	303	382	468	514	546	566	596	637
11		79	166	251	310	388	479	521	566	583	606	653
12	66	81	184	263	328	401	491	531	577	596	615	668
13		83	194	265	337	415	503	541	591	606	635	683
14		84	211	274	348	424	514	554	602	619	653	697
15		86	223	283	356	442	521	564	612	636	665	712
16	67	88	230	292	361	449	528	575	624	653	677	726
17		92	241	299	367	452	532	584	635	665	701	743
18		93	245	304	370	466	540	591	650	677	718	760
19	68	94	248	308	374	472	545	603	660	701	731	775
20		98	251	312	382	480	550	612	668	718	744	791
21		99	255	319	391	485	558	619	680	731	756	807
22	69	104	257	328	397	492	564	625	692	744	775	820
23		105	258	339	400	499	575	635	706	756	788	834
24		110	262	344	407	504	584	647	718	769	799	849
25		112	265	348	416	513	590	657	731	785	815	863
26		116	266	355	424	516	596	666	745	799	831	877
27		118	270	360	434	521	604	674	757	814	843	892
28		123	273	364	440	525	610	682	769	829	858	905
29	70	124	276	367	444	528	617	691	770	842	873	916

Table 7.2 (Continued)
Scaled Score to Percentile Rank Conversions

	Grade Placement											
PR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
30		128	279	370	448	532	624	699	785	857	882	927
31		130	283	374	452	540	631	708	802	867	892	937
32		133	286	375	456	545	636	717	809	878	905	946
33		136	289	383	464	550	642	724	822	887	916	955
34		138	292	385	469	557	648	730	835	898	918	967
35		140	298	391	472	558	657	739	842	908	927	979
36	71	142	301	395	477	564	665	750	849	918	946	990
37		148	305	398	481	567	674	757	858	927	957	1,001
38		149	308	402	487	570	681	770	871	936	967	1,012
39	72	155	312	407	491	576	688	780	881	942	973	1,021
40		160	316	412	494	581	698	790	890	952	981	1,033
41		166	319	417	497	586	705	802	897	961	990	1,047
42		172	323	424	499	590	712	809	903	980	1,019	1,062
43		179	326	428	506	596	720	817	910	990	1,034	1,076
44		188	331	434	510	604	727	825	916	1,004	1,050	1,092
45	73	197	336	440	514	610	733	834	926	1,013	1,060	1,107
46		208	339	444	516	613	742	843	934	1,020	1,070	1,122
47		217	342	445	521	619	749	851	944	1,034	1,093	1,133
48	74	223	345	450	525	625	757	859	949	1,041	1,100	1,144
49		232	348	453	528	631	769	871	958	1,050	1,104	1,147
50		239	351	456	532	636	778	883	967	1,063	1,113	1,152
51		241	355	460	538	643	783	889	980	1,070	1,122	1,158
52		245	357	464	544	650	789	894	990	1,084	1,136	1,169
53	75	246	360	468	548	657	797	901	1,004	1,094	1,146	1,178
54		248	364	472	553	665	806	906	1,013	1,104	1,152	1,189
55	76	250	367	477	557	674	813	910	1,028	1,112	1,155	1,193
56		252	369	481	558	677	823	917	1,039	1,122	1,164	1,199
57		254	372	485	560	681	834	924	1,051	1,136	1,171	1,206
58	77	255	375	490	564	688	841	938	1,060	1,137	1,181	1,215

Table 7.2 (Continued)
Scaled Score to Percentile Rank Conversions

	Grade Placement											
PR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
59		256	378	492	568	698	850	946	1,065	1,155	1,191	1,223
60	78	258	381	496	573	705	853	955	1,074	1,158	1,199	1,233
61		259	385	497	578	712	859	966	1,091	1,171	1,207	1,242
62	79	260	389	499	583	720	869	973	1,101	1,177	1,213	1,248
63		262	393	503	588	727	880	974	1,109	1,190	1,222	1,255
64	80	264	397	506	593	733	887	980	1,119	1,196	1,227	1,259
65		266	401	510	602	741	892	990	1,132	1,199	1,232	1,265
66	81	268	405	514	606	752	902	1,011	1,145	1,207	1,240	1,270
67		270	407	519	611	765	907	1,021	1,155	1,213	1,248	1,277
68		271	412	521	615	773	910	1,036	1,160	1,221	1,252	1,280
69	82	274	417	525	621	779	917	1,047	1,168	1,227	1,256	1,283
70		278	418	529	627	786	924	1,059	1,175	1,229	1,259	1,288
71	83	280	424	533	631	789	938	1,065	1,183	1,236	1,265	1,291
72	84	283	428	538	637	797	946	1,074	1,194	1,242	1,270	1,297
73	92	288	430	543	644	806	955	1,091	1,202	1,252	1,280	1,302
74	110	293	434	550	655	813	966	1,100	1,209	1,254	1,283	1,307
75	116	295	438	555	663	823	974	1,107	1,212	1,260	1,290	1,310
76	123	300	442	559	664	830	982	1,119	1,221	1,265	1,295	1,313
77	128	307	446	564	676	843	991	1,125	1,228	1,273	1,299	1,315
78	133	308	450	568	686	851	999	1,132	1,232	1,278	1,304	1,318
79	138	314	454	573	696	860	1,023	1,145	1,236	1,283	1,308	1,320
80	140	322	457	578	707	874	1,037	1,155	1,242	1,290	1,312	1,324
81		326	460	583	717	887	1,052	1,165	1,247	1,295	1,316	1,327
82	145	330	463	588	718	888	1,071	1,173	1,255	1,305	1,319	1,330
83	148	338	469	594	729	898	1,080	1,181	1,262	1,309	1,322	1,331
84	155	339	476	600	740	908	1,096	1,189	1,271	1,312	1,324	1,333
85	166	346	477	610	759	917	1,106	1,204	1,284	1,314	1,326	1,335
86	179	354	483	621	774	934	1,124	1,216	1,293	1,318	1,329	1,337
87	188	356	492	627	786	951	1,147	1,222	1,298	1,322	1,331	1,338

Table 7.2 (Continued)
Scaled Score to Percentile Rank Conversions

	Grade Placement											
PR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
88	208	367	496	631	792	965	1,165	1,227	1,306	1,326	1,335	1,341
89	225	369	500	642	801	974	1,173	1,232	1,310	1,329	1,336	1,342
90	239	377	508	657	821	999	1,181	1,244	1,314	1,333	1,338	1,343
91	246	389	513	679	843	1,032	1,192	1,254	1,318	1,334	1,340	1,344
92	250	392	519	686	851	1,037	1,210	1,271	1,322	1,336	1,342	1,345
93	254	401	525	704	868	1,067	1,224	1,289	1,326	1,338	1,344	
94	258	410	532	734	902	1,103	1,254	1,298	1,328	1,340	1,345	1,346
95	264	425	554	774	934	1,150	1,271	1,306	1,333	1,344		1,349
96	268	430	559	778	939	1,165	1,291	1,316	1,337	1,345	1,346	1,358
97	283	446	584	809	989	1,181	1,309	1,326	1,343	1,346		1,360
98	300	463	594	890	1,093	1,285	1,324	1,333	1,344		1,347	1,361
99	>338	>532	>778	>939	>1150	>1,309	>1,333	>1,340	>1,345	>1,347	>1,358	>1,368

Table 7.3
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 7.4
Normal Curve Equivalent to Percentile Rank Conversion

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

Table 7.4 (Continued)
Normal Curve Equivalent to Percentile Rank Conversion

NCE Range			NCE Range			NCE Range		
Low	High	PR	Low	High	PR	Low	High	PR
38.1	38.6	29	56.2	56.6	62	83.6	85.5	95
38.7	39.2	30	56.7	57.2	63	85.6	88.0	96
39.3	39.8	31	57.3	57.8	64	88.1	91.0	97
39.9	40.4	32	57.9	58.3	65	91.1	95.4	98
40.5	40.9	33	58.4	58.9	66	95.5	99.0	99

Table 7.5
Grade Equivalent to ZPD Conversions

GE	ZPD Range		GE	ZPD Range		GE	ZPD Range	
	Low	High		Low	High		Low	High
0.0	0.0	1.0	2.5	2.3	3.3	5.0	3.4	5.4
0.1	0.1	1.1	2.6	2.4	3.4	5.1	3.5	5.5
0.2	0.2	1.2	2.7	2.4	3.4	5.2	3.5	5.5
0.3	0.3	1.3	2.8	2.5	3.5	5.3	3.6	5.6
0.4	0.4	1.4	2.9	2.5	3.5	5.4	3.6	5.6
0.5	0.5	1.5	3.0	2.6	3.6	5.5	3.7	5.7
0.6	0.6	1.6	3.1	2.6	3.7	5.6	3.8	5.8
0.7	0.7	1.7	3.2	2.7	3.8	5.7	3.8	5.9
0.8	0.8	1.8	3.3	2.7	3.8	5.8	3.9	5.9
0.9	0.9	1.9	3.4	2.8	3.9	5.9	3.9	6.0
1.0	1.0	2.0	3.5	2.8	4.0	6.0	4.0	6.1
1.1	1.1	2.1	3.6	2.8	4.1	6.1	4.0	6.2
1.2	1.2	2.2	3.7	2.9	4.2	6.2	4.1	6.3
1.3	1.3	2.3	3.8	2.9	4.3	6.3	4.1	6.3
1.4	1.4	2.4	3.9	3.0	4.4	6.4	4.2	6.4
1.5	1.5	2.5	4.0	3.0	4.5	6.5	4.2	6.5
1.6	1.6	2.6	4.1	3.0	4.6	6.6	4.2	6.6
1.7	1.7	2.7	4.2	3.1	4.7	6.7	4.2	6.7
1.8	1.8	2.8	4.3	3.1	4.8	6.8	4.3	6.8
1.9	1.9	2.9	4.4	3.2	4.9	6.9	4.3	6.9
2.0	2.0	3.0	4.5	3.2	5.0	7.0	4.3	7.0
2.1	2.1	3.1	4.6	3.2	5.1	7.1	4.3	7.1
2.2	2.1	3.1	4.7	3.3	5.2	7.2	4.3	7.2
2.3	2.2	3.2	4.8	3.3	5.2	7.3	4.4	7.3
2.4	2.2	3.2	4.9	3.4	5.3	7.4	4.4	7.4

Table 7.5 (Continued)
Grade Equivalent to ZPD Conversions

GE	ZPD Range		GE	ZPD Range		GE	ZPD Range	
	Low	High		Low	High		Low	High
7.5	4.4	7.5	9.4	4.6	9.4	11.3	4.8	11.3
7.6	4.4	7.6	9.5	4.7	9.5	11.4	4.8	11.4
7.7	4.4	7.7	9.6	4.7	9.6	11.5	4.9	11.5
7.8	4.5	7.8	9.7	4.7	9.7	11.6	4.9	11.6
7.9	4.5	7.9	9.8	4.7	9.8	11.7	4.9	11.7
8.0	4.5	8.0	9.9	4.7	9.9	11.8	4.9	11.8
8.1	4.5	8.1	10.0	4.7	10.0	11.9	4.9	11.9
8.2	4.5	8.2	10.1	4.7	10.1	12.0	4.9	12.0
8.3	4.5	8.3	10.2	4.7	10.2	12.1	4.9	12.1
8.4	4.5	8.4	10.3	4.7	10.3	12.2	4.9	12.2
8.5	4.6	8.5	10.4	4.7	10.4	12.3	4.9	12.3
8.6	4.6	8.6	10.5	4.8	10.5	12.4	4.9	12.4
8.7	4.6	8.7	10.6	4.8	10.6	12.5	5.0	12.5
8.8	4.6	8.8	10.7	4.8	10.7	12.6	5.0	12.6
8.9	4.6	8.9	10.8	4.8	10.8	12.7	5.0	12.7
9.0	4.6	9.0	10.9	4.8	10.9	12.8	5.0	12.8
9.1	4.6	9.1	11.0	4.8	11.0	12.9	5.0	12.9
9.2	4.6	9.2	11.1	4.8	11.1	13.0	5.0	13.0
9.3	4.6	9.3	11.2	4.8	11.2			

Table 7.6
Scaled Score to Instructional Reading Level Conversions

Low	High	IRL
0	124	Pre-Primer (PP)
125	159	Primer (P)
160	167	1.0
168	176	1.1
177	185	1.2
186	194	1.3
195	200	1.4
201	212	1.5
213	220	1.6
221	229	1.7
230	238	1.8
239	247	1.9
248	256	2.0
257	266	2.1
267	274	2.2
275	284	2.3
285	293	2.4
294	304	2.5
305	315	2.6
316	324	2.7
325	336	2.8
337	345	2.9
346	359	3.0
360	369	3.1
370	379	3.2
380	394	3.3
395	407	3.4
408	423	3.5
424	439	3.6

Table 7.6 (Continued)
Scaled Score to Instructional Reading Level Conversions

Low	High	IRL
440	451	3.7
452	462	3.8
463	474	3.9
475	487	4.0
488	498	4.1
499	512	4.2
513	523	4.3
524	537	4.4
538	553	4.5
554	563	4.6
564	577	4.7
578	590	4.8
591	608	4.9
609	616	5.0
617	624	5.1
625	633	5.2
634	642	5.3
643	652	5.4
653	662	5.5
663	673	5.6
674	682	5.7
683	694	5.8
695	706	5.9
707	725	6.0
726	752	6.1
753	780	6.2
781	801	6.3
802	826	6.4
827	848	6.5

Table 7.6 (Continued)
Scaled Score to Instructional Reading Level Conversions

Low	High	IRL
849	868	6.6
869	877	6.7
878	904	6.8
905	916	6.9
917	918	7.0
919	920	7.1
921	922	7.2
923	924	7.3
925	927	7.4
928	930	7.5
931	933	7.6
934	936	7.7
937	939	7.8
940	942	7.9
943	947	8.0
948	953	8.1
954	960	8.2
961	966	8.3
967	970	8.4
971	974	8.5
975	983	8.6
984	988	8.7
989	998	8.8
999	1,011	8.9
1,012	1,022	9.0
1,023	1,034	9.1
1,035	1,042	9.2
1,043	1,050	9.3
1,051	1,058	9.4

Table 7.6 (Continued)
Scaled Score to Instructional Reading Level Conversions

Low	High	IRL
1,059	1,067	9.5
1,068	1,075	9.6
1,076	1,089	9.7
1,090	1,098	9.8
1,099	1,104	9.9
1,105	1,111	10.0
1,112	1,121	10.1
1,122	1,130	10.2
1,131	1,139	10.3
1,140	1,147	10.4
1,148	1,155	10.5
1,156	1,161	10.6
1,162	1,167	10.7
1,168	1,172	10.8
1,173	1,177	10.9
1,178	1,203	11.0
1,204	1,221	11.1
1,222	1,243	11.2
1,244	1,264	11.3
1,265	1,290	11.4
1,291	1,303	11.5
1,304	1,314	11.6
1,315	1,319	11.7
1,320	1,324	11.8
1,324	1,328	11.9
1,329	1,330	12.0
1,331	1,332	12.1
1,333	1,335	12.2
1,336	1,337	12.3

Table 7.6 (Continued)
Scaled Score to Instructional Reading Level Conversions

Low	High	IRL
1,338	1,340	12.4
1,341	1,341	12.5
1,342	1,342	12.6
1,343	1,343	12.7
1,344	1,344	12.8
1,345	1,345	12.9
1,346	1,400	Post-High School (PHS)

STAR Reading in the Classroom

Goal Setting for Student Progress Monitoring

By using STAR Reading on a regular basis, at least monthly or more often, 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 is appropriate for progress monitoring because it takes ten minutes or less to administer, it can be administered frequently, the results of the assessment can be graphed to show growth, and the assessment correlates well with the most widely selected state and standardized tests. The guidelines in this section will help teachers establish appropriate end-of-year goals consistent with No Child Left Behind.

Periodic Improvement

The Grade Equivalent Score is best used for measuring periodic improvement because it is reported in tenths of a grade. The correspondence between decimal value and month is shown in the table below.

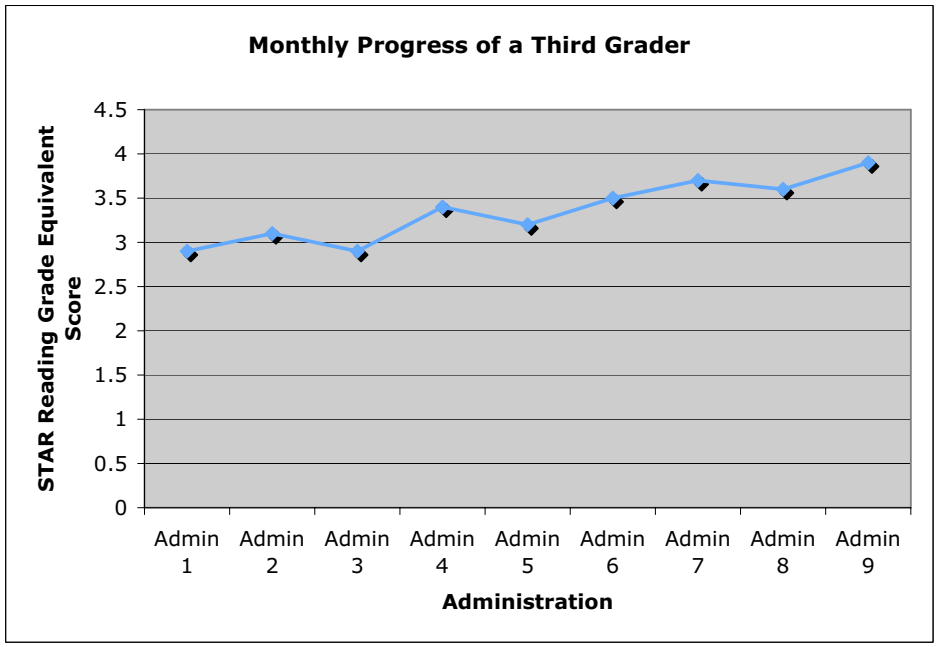
Table 8.1

Month	Decimal Equivalent
September	0.0
October	0.1
November	0.2
December	0.3
January	0.4
February	0.5
March	0.6
April	0.7
May	0.8
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 2 shows the score trajectory for a typical third-grade student for nine monthly administrations of STAR Reading. The student 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, S. Journal of Special Education, 37, 184-192)

Figure 2. Monthly Progress of a Third Grader



Adequate Yearly Progress

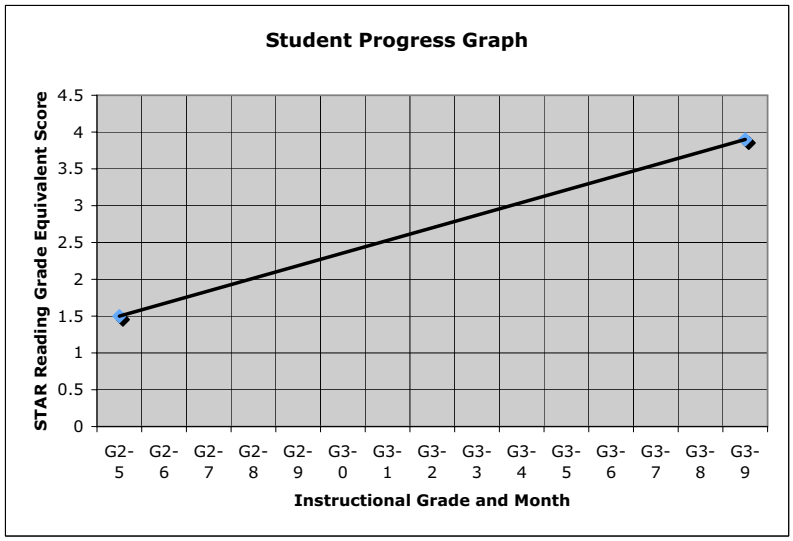
Establishing adequate yearly progress goals for students can also be accomplished using data from STAR Reading. If students are at or slightly below grade level, the Grade Equivalent Score may easily be used to measure adequate yearly progress within the current year. This score should be used when students are expected to reach grade equivalence within the same school year. In the example shown in Figure 2, the student is slightly below grade level at the beginning of the year. Choosing 3.9 as the end-of-year goal is appropriate because it is reasonable to expect that this goal will be reached within the year.

If students are significantly below grade level, creating a multi-year graph is necessary. The process requires several steps, but it is relatively straightforward.

- Use a recent STAR Reading Scaled Score as the baseline. A score from a single assessment may be used, but a more dependable baseline would be obtained by using the average of several recent administrations.
- Choose an appropriate end-of-year grade equivalent score with a realistic time frame, given the current literacy status of the student. For example, a student in the middle of second grade with a Grade Equivalent Score of 1.5 is unlikely to reach 2.9 by the end of the second grade. It is more likely that the student will reach grade equivalence by the end of the following year.

- Create a graph on which the x-axis represents the instructional time available and the y-axis represents Grade Equivalent Scores. Draw a progress line showing the student's current Grade Equivalent Score and the goal score at the end of the instructional period. Figure 3 shows the graph for the current example.

Figure 3.



- Note the end-of year Grade Equivalent Score needed to achieve the long-range goal. In the case of the example student, the goal for the end of second grade is approximately 2.3.
- On a monthly basis, plot the student's Grade Equivalent Score on the graph in order to ensure that the student is making adequate progress to achieve the intermediate end-of-year goal. Again, the student's progress might not be continuously upward. If the student's scores remain below the progress line for more than two consecutive months, it is advisable to reevaluate the instructional strategies being used with the student.

Instructional Planning with STAR Reading

The purpose of formative assessment 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 literacy status of students.

In many respects, STAR Reading is comparable to the oral fluency assessment traditionally used for progress monitoring. STAR Reading is sensitive to slight changes in reading skills, it has a high upper range so there is no ceiling effect for most grades, and it represents a skill that is critical for comprehending what is being read. The data generated by STAR Reading are as useful for instructional planning as are the results of a traditional oral fluency assessment

The "Guide to Reading Renaissance Goal-Setting Changes" 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 to use STAR Reading to:

- provide an accurate estimate of students' current reading status 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 genre that match their interests and challenge their abilities

Research Support

A number of research projects used STAR Reading to help teachers plan instruction. A study by Geoffrey Borman and Maritza 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[3]] 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. (pages 25-26)[4]

STAR Reading was also used as a planning and assessment tool in a study conducted by Sadusky and Brem (2002). Scores on the SAT-9 and STAR Reading were highly correlated (between 0.65 and 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.

---

3. 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.
4. Borman, G.D. and Dowling, N.M. *Testing the Reading Renaissance Program Theory: A Multilevel Analysis of Student and Classroom Effects on Reading Achievement.* University of Wisconsin-Madison, 2004.

> 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. (page 28)[5]

The Center for Research in Education Policy of the University of Memphis conducted a study of the School Renaissance mode. The report entitled "The Effect of School Renaissance on Student Achievement in Two Mississippi School Districts" is available on the University's Web site (http://crep.memphis.edu/web/research/). 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. (page 4)[6]

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

Researchers at the University of Georgia used STAR Reading to evaluate the effectiveness of the School Renaissance model in Georgia schools. The researches 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. (page 21)[7]

In an independent study of test scores from 1,100 predominantly Hispanic students in grades three through six, Louise 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 SAT 9 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 .69 to 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.[8]

---

5. Sadusky, L.A. and Brem, S.K. *The Integration of Renaissance Programs into an Urban Title I Elementary School, and Its Effect on School-wide Improvement*. Arizona State University, 2002.
6. Ross, S.M., Nunnery, J, and Goldfeder, E. *The Effect of School Renaissance on Student Achievement in Two Mississippi School Districts.* Center for Research in Educational Policy, University of Memphis, 2004. (http://crep.memphis.edu/web/research/)
7. Holmes, C.T., and Brown, C.L. *A Controlled Evaluation of a Total School Improvement Process, School Renaissance*. Technical Report. Athens, GA: University of Georgia, 2003. Available online: (http://www.coe.uga.edu/leadership/faculty/holmes/articles.html)
8. Bennicof-Nan, L. "A Correlation of Computer Adaptive, Norm Referenced, and Criterion Referenced Achievement Tests in Elementary Reading." Doctoral dissertation, The Boyer Graduate School of Education, Santa Ana, CA, 2002.

Growth Measurement

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 learning information system (LIS), 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 funds and for ensuring that those programs which will have the most educational impact on children are clearly identified.

Absolute Growth and Relative Growth

It is important to distinguish between two types of academic growth (or gains) that may be evidenced in test results. Absolute growth reflects any, and all, growth that has occurred. Relative growth reflects only growth that is above and beyond "normal" growth (i.e., beyond typical growth in a reference or norming group). 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 sample. In general, norm-referenced scores such as percentiles 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 if relative growth has occurred. That is, they are attempting to measure the impact of the intervention, or program, above and beyond normal growth.

The Pretest/Posttest Paradigm for Measuring Growth

The logical method for measuring growth (i.e., measuring effectiveness of educational interventions) is through the use of a pretest/posttest design. In such a design, each student is administered a test prior to the beginning of the intervention to establish a baseline measure. Then, each student is measured again at a later point in time (usually with a different, but equated, "form" of the same test) to see if the intervention is providing the desired outcome. The follow-up measurement may be at the end of the intervention, or may be done periodically throughout the course of the new program. Certainly, all of the issues relating to the adequacy of the test itself (e.g., in terms of core issues of reliability and validity) are applicable in order for this type of research to work properly. One key factor in conducting pretest/posttest designs is that if the same test is used both times, then the results may be compromised due to students having previously been

exposed to the test items. In an ideal situation, equivalent tests with no items in common should be administered. Subsequent administrations of a computerized adaptive test like the STAR Reading test are ideal for these types of assessments since this ensures that students get psychometrically parallel versions of the test at both pretest and posttest administrations, with no common items.

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' scale scores that could mask the true amount of growth.

Pretest/posttest with control group design

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

Pretest/posttest without a control group design

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. This has been the most commonly used method for measuring growth since it only involves a single group. This also allows you to apply the intervention to all students, without the need for a control group of any kind.

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 (7 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, we examine the difference between adjacent grade levels and, presuming even growth, interpolate between the empirical (observed) norms. Caution should be exercised when looking at growth which 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.

Using Scores to Measure Growth

There are a number of pieces of score information that are available from a standardized test such as the STAR Reading test. Among the scores available from the test are Scaled Scores, Percentile Ranks, Normal Curve Equivalents, and Grade Equivalents. All of these scores appear on the STAR Reading Growth Report. What are the differences between these scores, and which can best serve our purposes in attempting to measure growth?

Scaled Scores

Scaled Scores (SS) represent the student's score as expressed on a continuous vertical scale that spans all grade levels (1-12). The underlying vertical scale is derived as part of the test development process. In adaptive testing, students can receive different sets of items and still receive a comparable Scaled Score that represents their unique underlying ability level. Because Scaled Scores essentially map a student to a specific location on the underlying ability continuum, they can be useful in measuring absolute growth, and they are included in the STAR Reading 2.x and higher Growth Report.

Percentile Ranks

Percentiles, or more accurately Percentile Ranks, provide an easy way of relating a student's Scaled Score to the performance of a specified norm group (i.e., relating performance to one's peers). By providing a reference to a "standard" (i.e., norm group), norm referencing enhances the meaning of the test results. Percentiles range from 1 through 99 and define the percent of the norming sample that achieved lower Scaled Scores. Table 7.2 on page 59 lists the conversions from Scaled Scores to percentiles. In this table, the published Scaled Score to percentile table is abridged, showing only those values that are appropriate for administration of a test at the beginning of the school year. Internal to the STAR Reading program, the appropriate point-in-time Scaled Score to percentile relationship is determined dynamically as a normal part of the scoring process. Percentiles are probably the most common method for expressing results on norm-referenced tests because they are relatively easy to understand and to explain to others. If a Scaled Score of 445 corresponds to a percentile of 47 (at grade 4), it means that 47 percent of the norm group has Scaled Scores lower than 445. The main disadvantage of using percentiles is that they are not on an equal interval scale in terms of the underlying skill level. Quite simply, this means that gains of one percentile point, at various points along the scale, do not represent equal gains in achievement (skills). Because equal units on the percentile scale do not represent equal amounts of the underlying ability, it is not appropriate, or meaningful, to compute averages based on Percentile Ranks.

Normal Curve Equivalents

A transformation of the Percentile Ranks onto a scale that does have equal interval properties, the Normal Curve Equivalent (NCE) scale, must be performed before averaging of percentile information can be done. Although this conversion is accomplished automatically by STAR Reading software in the Growth Report, the tables used for converting percentiles to Normal Curve Equivalents (and from NCE to PR) are provided in Table 7.3 on page 63 and Table 7.4 on page 64. Thus, to obtain a measure of average performance, in

terms of Percentile Ranks, it is important to first convert them into NCE values using a conversion table (and thus transform to an equal interval scale) and then to compute the mean (average) of these NCEs. This is the process used for reporting group percentile information in STAR Reading software. Once the NCE scores have been averaged, it is allowable to "map" that mean (i.e., average) NCE back to its corresponding percentile point for reporting of growth.

Grade Equivalent Scores

Grade Equivalent scores represent one of the most commonly used methods for comparisons with a norm group. The GE scores reported for conventional tests can be misleading, however, because the GE may not reflect the student's performance on items of that grade level. For example, if a sixth-grader scores at grade level 4.1 on a traditional reading test, that does not mean that the sixth-grader is only capable of fourth-grade work. Technically, it means that the sixth-grader achieved a score that would be comparable with an "average" (defined as the 50th percentile) fourth-grade student after one month of instruction.

In an adaptive test like the STAR Reading test, the interpretation of Grade Equivalents more closely conforms to the common understanding, since the adaptive branching results in a test that is similar in both content and difficulty to what students at the grade-appropriate level would receive. It is important to note, however, that the amount of growth in terms of reading skill differs from grade to grade. Gains in reading are much greater in the early grades and tend to diminish with increasing grade level.

Grade Equivalents should not be used as the standard for growth per year or per grade. For a one-year period of time, the "normal" growth in GE scores would be 1.0 only at the 50th percentile. Below the 50th percentile, there is generally less than one year's growth (in terms of Grade Equivalents) during a one-year period. Above the 50th percentile, there will generally be more than one year's growth (in terms of Grade Equivalents) during a one-year period.

As with Percentile Ranks, Grade Equivalent scores should not be averaged in order to obtain the "typical" GE for the group since GEs are not on an equal-interval scale. Instead, the average of the group's Scaled Scores should be calculated, and the Grade Equivalent score corresponding to the average should be determined; Table 7.1 on page 54 may be used for this purpose.

Pretest/Posttest Studies of Growth Using a Single Sample Referenced Against Normative Data

The goal of this type of study is to determine if 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 if their 5th-grade posttest scores will result in a group percentile that is greater than 25.

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 U. S. 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.x test, the empirical norming period was in the month of April. The U. S. 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.

STAR Reading and the Elementary and Secondary Educational Act (ESEA, No Child Left Behind)

STAR Reading may be useful for districts and schools as they conform to the 2002 No Child Left Behind Legislation. For example, according to No Child Left Behind, starting in 2005, states must 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 1-12. 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.

Frequently Asked Questions

This chapter addresses a number of questions that educators have asked about STAR Reading tests and score interpretations.

What is the difference between criterion-referenced and norm-referenced testing?

The principal difference relates not so much to the tests themselves as to the interpretation of test scores. In criterion-referenced testing, the score is interpreted in terms of an unchanging criterion. Often the criterion-referenced test score interpretation addresses a performance-related question. For example: has the student attained mastery of specific curriculum objectives; does the student meet or exceed a specific performance standard; what proportion of the questions that measure knowledge of a specific content domain can the student answer correctly?

Norm-referenced score interpretations express the student's standing relative to the members of a specific reference group, such as his or her peers in the nationwide population of students in the same grade. A norm-referenced score is related to a changing criterion: changes in the definition of the reference group will be accompanied by changes in the norm-referenced scores. For example, when new norms are developed for a test, the distribution of test scores in the new reference group is typically different from the previous reference group to some extent, because of changes over time in both the reference population and the attribute the test measures. Such changes usually affect norm-referenced scores, including Percentile Ranks, Grade-Equivalent scores, and NCE scores.

Is the STAR Reading test criterion-referenced or norm-referenced?

The STAR Reading test was developed within a criterion-referenced framework, by writing test items designed to reflect reading ability that is characteristic of specific grade levels. However, the resulting tests have been subjected to a nationally representative norms development process. As a result, the STAR Reading test yields a variety of test scores, some of which (such as the Instructional Reading Level, or IRL) support criterion-referenced interpretations, and others of which (Percentile Ranks, Grade Equivalents, NCE scores) support norm-referenced interpretations. Teachers can use the STAR Reading criterion-referenced scores to estimate the student's level of functioning in reading, and its norm-referenced scores to assess students' standings relative to each other (based on Scaled Scores), to students in the same grade nationally (based on Percentile Ranks), and to students in other grades (based on Grade Equivalent scores). "Score Definitions" (page 31) describes all of the scores reported by STAR Reading 2.x and higher software; it also provides information on how to interpret each one.

Why is it that GE and IRL scores sometimes differ?

These two scores are both expressed in terms of grade levels, but that is the only similarity between them. The Grade Equivalent (GE) is a norm-referenced score that indicates the grade level at which the performance of students in the normative population was most similar to that of the student who is the subject of interest. For example, if a student received a GE score of 4.7, this means that this student's Scaled

Score is equivalent to the median Scaled Score of students in the normative population who were in the 7th month of the 4th grade — half of those students scored higher, half scored lower.

In contrast, the Instructional Reading Level (IRL) is a criterion-referenced score that estimates the grade level of written material at which the student can most effectively be taught. It is derived by using the Scaled Score to estimate the student's proficiency on each of 14 sets of test items that have been graded from K to 13 according to the EDL Core Vocabulary listing. The IRL is defined as the highest of those 14 grade levels at which the student can answer 80% or more of the test items correctly.

GE and IRL scores are highly correlated, but they do not connote the same thing, and there is no reason to expect them to be identical. While they may coincide in some cases, they can differ markedly, particularly for students who are significantly above or below average, where the functional reading level is likely to differ substantially from the grade level.

How do ZPD ranges fit in?

The Zone of Proximal Development defines the reading level range from which the student should be selecting books for reading practice in order to achieve optimal growth in literacy skills.

The ZPD is derived from a student's demonstrated Grade Equivalent score. 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 *Research Foundation for Reading Renaissance Goal Setting Practices* (2003), which was published by Renaissance Learning. This information is also distributed at Renaissance Learning seminars. Table 7.5 on page 66 contains the relationship between GE and ZPD.

Do all students receive longer authentic text passage items for the last five questions of the STAR Reading 2.x and higher tests?

No. Only third through twelfth graders receive authentic text passage items. These new items were extracted from passages of authentic fiction and nonfiction text. Passages at the third grade level are about 30 words long, while passages at the high school level are about 100 words long. First and second graders receive shorter vocabulary-in-context items throughout the STAR Reading 2.x and higher tests.

How can the STAR Reading test determine a child's reading level in less than ten minutes?

Short test times are possible because the STAR Reading test is computer-adaptive. It adapts to test a student at his or her level of proficiency. Because the STAR Reading test can adapt and adjust to the student with virtually every question, it is more efficient than conventional pencil and paper tests, and acquires more information about a student's reading ability in less time. This means the STAR Reading test can achieve measurement precision comparable to a conventional test that takes two or more times as long to administer.

How are STAR Reading 2.x and higher Diagnostic Reports constructed from the test results?

The contents of Diagnostic Reports are based on diagnostic codes (though the codes themselves are not printed on the reports). Diagnostic codes, in turn, are based on two factors. The first is the Grade Equivalent (GE) score that the student achieved on the STAR Reading test. The second is the Percentile Rank (PR) that the student achieved on the same test. The resulting diagnostic code determines which descriptive text and prescriptive recommendations appear on the Diagnostic Report for a student. Diagnostic Reports for students performing below the 25th percentile include additional prescriptive information helpful for assisting those students.

The diagnostic codes and accompanying descriptive text were developed by a reading specialist using standard scope and sequence paradigms from the field of Reading Education. STAR Reading 2.x and higher RP Diagnostic Reports therefore contain generalized descriptions of reading skills and development patterns that are based on the Grade Equivalent and Percentile Rank score. They also contain prescriptive information to encourage and promote optimal growth in reading.

How does the STAR Reading test compare to other standardized tests?

Very well. The STAR Reading test has a standard error and reliability that are very comparable to other standardized norm-referenced tests. Also, STAR Reading test results correlate well with results from these other test instruments. When performing our national norming of the STAR Reading 2.x test, we also gathered student performance data from several other commonly used reading tests. These data comprised more than 12,000 student test results from test instruments including CAT, ITBS, MAT, Stanford, TAAS, CTBS, and others. We computed correlation coefficients between STAR Reading 2.x results and results of each of these test instruments for which we had sufficient data. These correlation coefficients are included in “Reliability and Validity” (pages 44-53). Using IRT computer-adaptive technology, the STAR Reading test achieves its results with fewer test items and shorter test times than other standardized norm-referenced tests.

What are some of the other standardized tests that might be compared to the STAR Reading test?

CAT - California Achievement Test (K to 12)

Designed to measure achievement in the basic skills commonly found in state and district curricula.

CTBS - Comprehensive Test of Basic Skills (K to 12)

Modular testing system that evaluates students’ academic achievement from K-12. It measures the basic content areas — reading, language, spelling, mathematics, study skills, science, and social studies.

Gates -MacGinitie Reading Test (K to 12)

Designed to assess student achievement in reading.

ITBS - Iowa Test of Basic Skills (K to 9)

Designed to provide for comprehensive and continuous measurement of growth in the fundamental skills, vocabulary, reading, the mechanics of writing, methods of study, and mathematics.

MAT - Metropolitan Achievement Test (1 to 12)

Designed to measure the achievement of students in the major skill and content areas of the school curriculum.

Stanford Achievement Test (1 to 12)

Designed to measure the important learning outcomes of the school curriculum. Measures student achievement in reading, mathematics, language, spelling, study skills, science, social studies, and listening.

TAKS - Texas Assessment of Knowledge and Skills (3 to 11)

Statewide Texas Education Agency mandated criterion-referenced test used to assess student and school system performance. Includes tests in reading, math, writing, science, and social studies. Passage of a grade 10 exit exam is required for high school graduation.

Why do some of my students who took STAR Reading tests have scores that are widely varying from the results of our other standardized test program?

The simple answer to this is that it is more than likely the result of the Standard Error of Measurement (SEM) of both testing instruments. Unfortunately, such a simple answer hides the complexity of the many factors that contribute to measurement errors inherent in psychometric instruments. You will find that, on the whole, STAR Reading results will agree very well with almost all of the other standardized reading test results.

All standardized test scores have measurement error. The STAR Reading measurement error is comparable to most other standardized tests. When one compares the results from different tests taken at different times, it is not unusual to see differences in test scores ranging from two to five grade levels. This is true when comparing results from other test instruments as well. Standardized tests provide approximate measurements. The STAR Reading test is no different in this regard, but its adaptive nature makes its scores more reliable than conventional test scores near the minimum and maximum scores on a given form. A common shortcoming of conventional tests involves "floor" and "ceiling" effects at each test level. The STAR Reading test is not subject to this shortcoming because of its adaptive branching and large item bank.

Other factors, such as student motivation and the testing environment, are also different for STAR Reading and high-stakes tests.

Why do we see a significant number of our students performing at a lower level now than they were nine weeks ago?

This is a result of measurement error. As mentioned just above, all psychometric instruments, including the STAR Reading test, have some level of measurement error associated with them. The "Reliability and Validity" chapter discusses standard error of measurement (SEM) in depth (beginning on page 42); it should be referred to in order to better understand this issue.

The standard error of the average results for a group is substantially lower than it is for individual test scores. Therefore, more frequent testing to measure the progress of classes, grades or school populations will be less susceptible to measurement error.

How many items will a student be presented with when taking a STAR Reading test?

The STAR Reading 2.x and higher RP tests administer the same number of items — 25 — to all students. Students tested at grade levels 3 through 12 receive 20 vocabulary-in-context items and five authentic text passage items. For students tested at grade levels 1 and 2, all 25 items are vocabulary-in-context items.

How many items does the STAR Reading test have at each grade level?

The STAR Reading test has enough items at each grade level that students can be tested five times per year and should not be presented with the same material they have already been tested on in the same year. Generally, the STAR Reading software will not administer the same item twice to a student within a six-month period.

What guidelines are offered as to whether a student can be tested using STAR Reading software?

In general the student should have a reading vocabulary of at least 100 words. In other words, the student should have at least beginning reading skills. Practically, if the student can work through the practice questions unassisted, that student should be able to be tested using STAR Reading software. If the student has a lot of trouble getting through the practice, it is likely that he or she does not possess the basic skills necessary to be measured by STAR Reading software.

How will students with a fear of taking tests do with STAR Reading tests?

Students who have a fear of tests should be less disadvantaged by the STAR Reading test than they are in the case of conventional tests. The STAR Reading test purposely starts out at a level that most students will find to be very easy. This was done in order to give almost all students immediate success with the STAR Reading test. Once the student has had an opportunity to gain some confidence with the relatively easy material, the STAR Reading test moves into more challenging material in order to assess the level of reading proficiency.

In addition, most students find it fun to take STAR Reading tests on the computer, which helps relieve some test anxiety.

Is there any way for a teacher to see exactly which items a student answered correctly and which he or she answered incorrectly?

No. This was done for two reasons. First, in computer-adaptive testing, the student's performance on individual items is not as meaningful as the pattern of responses to the entire test. The student's pattern of performance on all items taken together forms the basis of the scores STAR Reading reports. Second, for purposes of test security, we decided to do everything possible to protect our items from compromise and overexposure.

How did you choose schools to participate in the norming of STAR Reading 2.x?

Schools were chosen to be representative of the nation as a whole. The sample was balanced by geographic region, district size, socioeconomic status, and public vs. private funding characteristics. In all, the norming sample included 269 schools and 29,627 students in grades 1 through 12. The final norms tables were weighted to account for deviations in the characteristics of the sample compared to the national norms.

What evidence do we have that STAR Reading software will perform as claimed?

This evidence comes in two forms. First, we have demonstrated test-retest reliability estimates that are very good. Second, the correlation of STAR Reading 2.x results with those of other standardized tests is also quite impressive. See "Reliability and Validity" (starting on page 39) for reliability and validity data.

Can or should the STAR Reading test replace a school's current standardized tests?

This is up to the school system to decide, although this is not what the STAR Reading test was primarily designed to do. The primary purpose of the STAR Reading test is to provide teachers with a tool to improve the instructional match for each student. Every school system has to consider its needs in the area of reading assessment and make decisions as to what instruments will meet those needs. We are happy to provide as much information as we can to help schools make these decisions, but we cannot make the decision for them.

What is Item Response Theory?

Item Response Theory (IRT) is an approach to psychometric test design and analysis that uses mathematical models that describe what happens when an examinee is administered a particular test question. IRT models give the probability of answering an item correctly as a function of the item's difficulty and the examinee's ability. More information can be found in any text on modern test theory.

What are the Cloze and Maze procedures?

These are terms for different kinds of fill-in-the-blank exercises that test a student's ability to create meaning from contextual information, and as such have elements in common with the STAR Reading 2.x and higher test design.

Appendix A: List of Participating Schools

School Name	City	State	Region
Chester Valley Elementary School	Anchorage	AK	W
Chiniak School	Chiniak	AK	W
Whittier School	Whittier	AK	W
Academy for Science and Foreign Language	Huntsville	AL	SE
Brookwood Forest Elementary School	Birmingham	AL	SE
Brookwood High School	Brookwood	AL	SE
Central Elementary School	Tuscaloosa	AL	SE
Cleburne County High School	Heflin	AL	SE
Craighead Elementary School	Mobile	AL	SE
Davis Elementary School	Theodore	AL	SE
Erwin High School	Birmingham	AL	SE
Evans Elementary School	Albertville	AL	SE
Forest Avenue Elementary School	Montgomery	AL	SE
Kinston School	Kinston	AL	SE
Muscle Shoals Middle School	Muscle Shoals	AL	SE
New Brockton Elementary School	New Brockton	AL	SE
Princeton Alternative Elementary School	Birmingham	AL	SE
Ranburne High School	Ranburne	AL	SE
Reeltown Elementary and High School	Notasulga	AL	SE
Smith Elementary School	Gadsden	AL	SE
Wilson School	Florence	AL	SE
Bald Knob Middle School	Bald Knob	AR	SE
Green County Technical Intermediate School	Paragould	AR	SE
Izard County School	Brockwell	AR	SE
Viola Public School	Viola	AR	SE
Viola Public School District	Viola	AR	SE
Carson Junior High School	Mesa	AZ	W
Fowler Elementary School	Phoenix	AZ	W

School Name	City	State	Region
Kachina Elementary School	Glendale	AZ	W
Litchfield Elementary School	Litchfield Park	AZ	W
Scott Libby Elementary School	Litchfield Park	AZ	W
Sopori Elementary School	Amado	AZ	W
Clairemont Senior High School	San Diego	CA	W
Eagle Ranch Elementary School	Victorville	CA	W
Los Molinos Elementary School	Los Molinos	CA	W
Lower Lake High School	Lower Lake	CA	W
Rancho Cucamonga Middle School	Rancho Cucamonga	CA	W
Rio Canyon School	Reedley	CA	W
San Gabriel Mission Elementary School	San Gabriel	CA	W
Silva Valley Elementary School	El Dorado Hills	CA	W
Solano Christian Academy	Fairfield	CA	W
Southwest Junior High School	San Diego	CA	W
SS Simon and Jude School	Huntington Beach	CA	W
St. Bernadette School	Los Angeles	CA	W
St. Bonaventure School	Huntington Beach	CA	W
St. Stephen Lutheran School	Fallbrook	CA	W
Woodridge Elementary School	Sacramento	CA	W
Yerba Buena High School	San Jose	CA	W
Academy of Charter Schools	Denver	CO	W
Bricker Elementary School	Colorado Springs	CO	W
Force Elementary School	Denver	CO	W
Gateway Elementary School	Woodland Park	CO	W
Liberty High School	Colorado Springs	CO	W
Parkview Elementary School	Rangely	CO	W
Silverthorne Elementary School	Silverthorne	CO	W
Soda Creek Elementary School	Steamboat Springs	CO	W
Willow Creek Elementary School	Englewood	CO	W
Edgerton Elementary School	New London	CT	NE
Franklin Elementary School	Meriden	CT	NE

School Name	City	State	Region
Hill Central Elementary School	New Haven	CT	NE
Sage Park Middle School	Windsor	CT	NE
Fairview Elementary School	Dover	DE	NE
Harbour View Elementary School	Summerfield	FL	SE
Heights Elementary School	Fort Myers	FL	SE
LeHigh Acres Middle School	Lehigh Acres	FL	SE
Lipscomb Elementary School	Pensacola	FL	SE
Malone High School	Malone	FL	SE
Orange Park Junior High School	Orange Park	FL	SE
Rolling Hills Elementary School	Orlando	FL	SE
Ruskin Elementary School	Ruskin	FL	SE
Three Oaks Elementary School	Fort Myers	FL	SE
Baker Elementary School	Acworth	GA	SE
Brown Elementary School	Jonesboro	GA	SE
Ephesus Elementary School	Roopville	GA	SE
Fayette County High School	Fayetteville	GA	SE
Jones-Wheat Elementary School	Bainbridge	GA	SE
McNiar High School	Atlanta	GA	SE
Norman Park Elementary School	Norman Park	GA	SE
Oakcliff Elementary School	Atlanta	GA	SE
Pine Mountain Middle School	Kennesaw	GA	SE
Summerville Elementary School	Summerville	GA	SE
Tattnall Square Academy	Macon	GA	SE
Hilo High School	Hilo	HI	W
James B. Castle High School	Kaneohe	HI	W
Pa'Auilo Elementary-Intermediate School	Paauilo	HI	W
Deep River-Millersburg Elementary School	Millersburg	IA	MW
Gehlen Catholic School	Le Mars	IA	MW
Longfellow Elementary School	Des Moines	IA	MW
Paul Norton Elementary School	Bettendorf	IA	MW
Reinbeck Elementary School	Reinbeck	IA	MW

School Name	City	State	Region
River Valley Elementary School	Cushing	IA	MW
Sacred Heart Elementary School	Boone	IA	MW
Springville Elementary School	Springville	IA	MW
Cynthia Mann Elementary School	Boise	ID	W
Gate City Elementary School	Pocatello	ID	W
Les Bois Junior High School	Boise	ID	W
Midway Elementary School	Menan	ID	W
Raft River Elementary-Junior High School	Malta	ID	W
St. Mark's Elementary School	Boise	ID	W
Amboy Central Elementary School	Amboy	IL	MW
Chase Elementary School	Chicago	IL	MW
Dunbar Elementary School	E. Saint Louis	IL	MW
East High School	Rockford	IL	MW
Elgin High School	Elgin	IL	MW
Farragut Career Academy High School	Chicago	IL	MW
Forsyth Grade School	Forsyth	IL	MW
Frankfort Square Elementary School	Frankfort	IL	MW
Fulton Elementary School	Tinley Park	IL	MW
Gallistel Language Academy	Chicago	IL	MW
Hayt Elementary School	Chicago	IL	MW
Jefferson Elementary School	Dixon	IL	MW
Lenart Regional Gifted Center	Chicago	IL	MW
Lincoln Elementary School	Dixon	IL	MW
Madison Elementary School	South Holland	IL	MW
Maroa-Forsyth Junior High School	Maroa	IL	MW
Martinsville Elementary School	Martinsville	IL	MW
Monroe Elementary School	Chicago	IL	MW
Murphy Elementary School	Chicago	IL	MW
Oriole Park Elementary School	Chicago	IL	MW
Prairieview Elementary School	Bartlett	IL	MW
Silver Street Elementary School	Olney	IL	MW

School Name	City	State	Region
Sparland Elementary School	Sparland	IL	MW
St. Dorothy Elementary School	Chicago	IL	MW
Stevenson Elementary School	Chicago	IL	MW
V. Blanche Graham Elementary School	Naperville	IL	MW
Virginia Lake Elementary School	Palatine	IL	MW
Elm Road Elementary School	Mishawaka	IN	MW
Fairview Elementary School	Richmond	IN	MW
La Salle High School	South Bend	IN	MW
Lincoln Elementary School	Columbus	IN	MW
Michigan High School	Michigan City	IN	MW
Muessel Elementary School	South Bend	IN	MW
Orchard Park Elementary School	Indianapolis	IN	MW
Park Elementary School	Michigan City	IN	MW
Springville Elementary School	Springville	IN	MW
Twin Branch Elementary School	Mishawaka	IN	MW
Wison Elementary School	Hammond	IN	MW
Wison Elementary School	Jeffersonville	IN	MW
Alta Brown Elementary School	Garden City	KS	MW
Bishop Elementary School	Topeka	KS	MW
Edith Scheuerman Elementary School	Garden City	KS	MW
Faris Elementary School	Hutchinson	KS	MW
Garden Plain Elementary School	Garden Plain	KS	MW
Garfield Elementary School	Liberal	KS	MW
Georgia Matthews Elementary School	Garden City	KS	MW
Girard High School	Girard	KS	MW
Girard Middle School	Girard	KS	MW
Haderlein Elementary School	Girard	KS	MW
Haysville Middle School	Haysville	KS	MW
Highland Elementary School	Highland	KS	MW
Jefferson West Elementary School	Meriden	KS	MW
Jennie Barker Elementary School	Garden City	KS	MW

School Name	City	State	Region
Natoma High School	Natoma	KS	MW
Plains Elementary School	Plains	KS	MW
Robinson Elementary School	Augusta	KS	MW
South Vernon Elementary School	Winfield	KS	MW
St. Joseph Elementary School	Conway Springs	KS	MW
Washington Elementary School	Liberal	KS	MW
Wilbur Middle School	Wichita	KS	MW
Apollo High School	Owensboro	KY	SE
Beech Fork Elementary School	Helton	KY	SE
Eastern Elementary School	Georgetown	KY	SE
Elkhorn Middle School	Frankfort	KY	SE
F. T. Burns Middle School	Owensboro	KY	SE
Larue County High School	Hodgenville	KY	SE
Northern Elementary School	Georgetown	KY	SE
Philpot Elementary School	Philpot	KY	SE
Ponderosa Elementary School	Catlettsburg	KY	SE
Shepherd Elementary School	Columbia	KY	SE
Yealey Elementary School	Florence	KY	SE
Alleman Middle School	Lafayette	LA	SE
Benton Elementary School	Benton	LA	SE
Ballard Elementary School	Saugus	MA	NE
Burke Elementary School	Peabody	MA	NE
Elm Street Elementary School	Gardner	MA	NE
Essex Christian Academy	Beverly	MA	NE
First Lutheran School	Holyoke	MA	NE
Florence Sawyer School	Bolton	MA	NE
High School of Commerce	Springfield	MA	NE
Aberdeen Middle School	Aberdeen	MD	NE
Berlin Intermediate School	Berlin	MD	NE
Braddock Middle School	Cumberland	MD	NE
Ecole Secondaire Casavant	St. Hyacinthe	MD	NE

School Name	City	State	Region
Hillcrest Elementary School	Frederick	MD	NE
Salem Avenue Elementary School	Hagerstown	MD	NE
Spencerville Adventist Academy	Silver Spring	MD	NE
Edward Little High School	Auburn	ME	NE
Lewiston High School	Lewiston	ME	NE
Littleton Elementary School	Monticello	ME	NE
Oxford Hills Middle School	South Paris	ME	NE
Southside Elementary School	Houlton	ME	NE
Adams Elementary School	Waterford	MI	MW
Bedford Senior High School	Temperance	MI	MW
Hamilton Parsons Elementary School	Leonard	MI	MW
Harms Elementary School	Detroit	MI	MW
Kempton Elementary School	Saginaw	MI	MW
L. E. White Middle School	Allegan	MI	MW
Lincoln Elementary School	South Haven	MI	MW
Lincoln Elementary School	Roseville	MI	MW
Linsday Elementary School	Bay City	MI	MW
Milwood Middle School	Kalamazoo	MI	MW
Willow Elementary School	Lansing	MI	MW
Albany Elementary School	Albany	MN	MW
Armatage Elementary School	Minneapolis	MN	MW
Johnsonville School	Blaine	MN	MW
Lakeview Elementary School	Albert Lea	MN	MW
Solway Elementary School	Bemidji	MN	MW
St. Mary New Monmout	Middletown	MN	MW
Carman Trails Elementary School	Ballwin	MO	MW
Central Computers Greek High School	Kansas City	MO	MW
Crestwood Elementary School	Saint Louis	MO	MW
Flynn Park Elementary School	Saint Louis	MO	MW
Foreign Language K-8 School	Kansas City	MO	MW
Gideon Elementary School	Gideon	MO	MW

School Name	City	State	Region
Hurley Elementary School	Hurley	MO	MW
Hurley High School	Hurley	MO	MW
Kelsey Norman Elementary School	Joplin	MO	MW
Lesterville Ranch Campus	Black	MO	MW
Meservey Elementary School	Kansas City	MO	MW
O'Neal Elementary School	Poplar Bluff	MO	MW
Pickett Elementary School	Saint Joseph	MO	MW
Plato School	Plato	MO	MW
Rocky Comfort Elementary School	Rocky Comfort	MO	MW
Sheldon Elementary School	Sheldon	MO	MW
Strafford Elementary School	Strafford	MO	MW
Gentry High School	Indianola	MS	SE
Winston Academy	Louisville	MS	SE
Boomfield Elementary School	Bloomfield	MT	W
Cut Bank Middle School	Cut Bank	MT	W
Grantsdale Elementary School	Hamilton	MT	W
Cameron Park Elementary School	Hillsborough	NC	SE
Chadbourn Middle School	Chadbourn	NC	SE
Chocowinity High School	Chocowinity	NC	SE
Contentnea Elementary School	Kinston	NC	SE
Eastern Guilford High School	Gibsonville	NC	SE
Glade Creek Elementary School	Ennice	NC	SE
North Mecklenburg High School	Huntersville	NC	SE
Ocracoke School	Ocracoke	NC	SE
Our Lady of Lourdes School	Raleigh	NC	SE
Randolph Elementary School	Asheville	NC	SE
Vanceboro Farm Life Elementary School	Vanceboro	NC	SE
North Central Public School	Rocklake	ND	MW
Selfridge High School	Selfridge	ND	MW
Shiloh Christian School	Bismarck	ND	MW
Simle Middle School	Bismarck	ND	MW

School Name	City	State	Region
Ansley Public School	Ansley	NE	MW
Crete Elementary School	Crete	NE	MW
Crete Junior-Senior High School	Crete	NE	MW
La Vista Junior High School	La Vista	NE	MW
McDonald Elementary School	North Platte	NE	MW
North Park Elementary School	Broken Bow	NE	MW
Roosevelt Elementary School	Scottsbluff	NE	MW
Wolback Public School	Wolback	NE	MW
Unity Elementary School	Newport	NH	NE
Eastern Christian Middle School	Wyckoff	NJ	NE
Joseph F. Brandt Elementary/Middle School	Hoboken	NJ	NE
Monongahela Middle School	Sewell	NJ	NE
Oakcrest High School	Mays Landing	NJ	NE
Perth Amboy Elementary School #10	Perth Amboy	NJ	NE
Mesilla Park Elementary School	Las Cruces	NM	W
Riverside Elementary School	Sunland Park	NM	W
Becker Middle School	Las Vegas	NV	W
Jackpot Combined School	Jackpot	NV	W
Lund Elementary School	Lund	NV	W
Lyon Middle School	Overton	NV	W
McGill Elementary School	McGill	NV	W
Abraham Wing Elementary School	Glens Falls	NY	NE
Bemus Elementary School	Bemus Point	NY	NE
East Middle School	Brentwood	NY	NE
Floyd Bell Elementary School	Kirkwood	NY	NE
Franklyn Barry Elementary School	Cortland	NY	NE
Greece Christian School	Rochester	NY	NE
I S 061 Atwell School	Brooklyn	NY	NE
Kernan Elementary School	Utica	NY	NE
Northstar Christian Academy	Rochester	NY	NE
P S 115 Glen Oaks School	Floral Park	NY	NE

School Name	City	State	Region
Pine Brook Elementary School	Rochester	NY	NE
Randall Elementary School	Cortland	NY	NE
Regina Coeli School	Hyde Park	NY	NE
Ripley Central School	Ripley	NY	NE
Sacred Heart Elementary School	Utica	NY	NE
St. Ann Elementary School	Hornell	NY	NE
St. Brigid School	Brooklyn	NY	NE
St. Peter Lutheran School	Lockport	NY	NE
Albert Hart Middle School	Cleveland	OH	MW
Brantner Lane Elementary School	Cincinnati	OH	MW
Christ The King School	Akron	OH	MW
Christian Academy	Sidney	OH	MW
East Liverpool Christian School	East Liverpool	OH	MW
Grover Cleveland Middle School	Zanesville	OH	MW
Schumacher Elementary School	Akron	OH	MW
Walker Elementary School	Canton	OH	MW
Warren Christian School	Warren	OH	MW
Broadmoore Elementary School	Oklahoma City	OK	W
Carrier Elementary School	Guymon	OK	W
Independence Middle School	Yukon	OK	W
Justus-Tiawah School South	Claremore	OK	W
Northeast Elementary School	Guymon	OK	W
Parkland Elementary School	Yukon	OK	W
Purcell Elementary School	Purcell	OK	W
Roosevelt Elementary School	Altus	OK	W
Salyer Elementary School	Guymon	OK	W
Sooner Rose Elementary School	Midwest City	OK	W
Washington Elementary School	Ponca City	OK	W
Whittier Elementary School	Lawton	OK	W
Woodlawn Elementary School	Sapulpa	OK	W
Judson Middle School	Salem	OR	W

School Name	City	State	Region
Aleppo Elementary School	Aleppo	PA	NE
Altoona Central Catholic Middle School	Altoona	PA	NE
Cambria County Christian School	Johnstown	PA	NE
Central Dauphin High School	Harrisburg	PA	NE
Hope Lutheran School	Levittown	PA	NE
J S Jenks Elementary School	Philadelphia	PA	NE
Market Street Elementary School	Warren	PA	NE
Mayfair Elementary School	Philadelphia	PA	NE
Northside Elementary School	Mechanicsburg	PA	NE
Riverside JR/SR High School	Taylor	PA	NE
Southside Middle School	Hookstown	PA	NE
St Boniface Elementary School	Kersey	PA	NE
St Joseph Elementary	Warren	PA	NE
West End Elementary School	Meadville	PA	NE
Aiken Middle School	Aiken	SC	SE
Barwell Christian School	Blackville	SC	SE
Bowman Middle/High School	Bowman	SC	SE
C E Murray High School	Greeleyville	SC	SE
Gadsen Elementary School	Gadsen	SC	SE
Jackson Middle School	Jackson	SC	SE
Simpson Academy	Easley	SC	SE
Slater-Marietta Elementary School	Marietta	SC	SE
St Joseph's School	Columbia	SC	SE
Lyman Middle School	Kennebec	SD	MW
South Middle School	Rapid City	SD	MW
Williams Elementary School	Mitchell	SD	MW
Woodrow Wilson Elementary School	Rapid City	SD	MW
Adrian Burnett Elementary School	Knoxville	TN	SE
Andrew Jackson Elementary School	Jackson	TN	SE
Black Fox Elementary School	Cleveland	TN	SE
Blythe Avenue Elementary School	Cleveland	TN	SE

School Name	City	State	Region
McEwen Elementary School	McEwen	TN	SE
Middle Tennessee Christian School	Murfreesboro	TN	SE
Milan Middle School	Milan	TN	SE
Moore Elementary School	Franklin	TN	SE
Riverdale Elementary School	Germantown	TN	SE
Riverside High School	Parsons	TN	SE
Stony Fork School	Caryville	TN	SE
Sullivan Elementary School	Kinsport	TN	SE
Thomas Intermediate School	Shelbyville	TN	SE
Trousdale County Elementary School	Hartsville	TN	SE
Unity Elementary/JR High School	Petersburg	TN	SE
Westwood Elementary School	Manchester	TN	SE
White Pine Elementary School	White Pine	TN	SE
Anderson Elementary School	Conroe	TX	W
Bastrop High School	Bastrop	TX	W
Bishop T Gorman Middle School	Tyler	TX	W
C Martinez Elementary School	Houston	TX	W
Canyon Junior High School	Canyon	TX	W
Childress High School	Childress	TX	W
Christ the King School	Lubbock	TX	W
Community Christian School	Orange	TX	W
Cooper School	Cooper	TX	W
Crawford Elementary School	Crawford	TX	W
Dallas Christian School	Mesquite	TX	W
Desert View Middle School	El Paso	TX	W
Dorsey Elementary School	Rowlett	TX	W
Dublin JR High School	Dublin	TX	W
East Texas Christian Academy	Tyler	TX	W
Eastwood Knolls Elementary School	El Paso	TX	W
Edgemere Elementary School	El Paso	TX	W
Elysian Fields Middle School	Elysian Fields	TX	W

School Name	City	State	Region
Hedley School	Hedley	TX	W
Heflin Elementary School	Houston	TX	W
Hillcrest Elementary School	Plainview	TX	W
H O Whitehurst Elementary School	Groesbeck	TX	W
John F Kennedy Elementary School	Corpus Cristi	TX	W
Lamarque High School	Lamarque	TX	W
Lee Elementary School	Marshall	TX	W
Lorena Elementary School	Lorena	TX	W
Lorena High School	Lorena	TX	W
Lovett Ledger Intermediate School	Copperas Cove	TX	W
Mission West Elementary School	Houston	TX	W
Montclair Elementary School	Garland	TX	W
New Caney Elementary School	New Caney	TX	W
Northampton Elementary School	Spring	TX	W
Olton High School	Olton	TX	W
Olton Junior High School	Olton	TX	W
Our Lady of Perpetual Help School	Selma	TX	W
Our Savior Lutheran School	Houston	TX	W
Roosvelt Wilson Elementary School	Texas City	TX	W
Southwest Christian School	Fort Worth	TX	W
St John's Episcopal School	Abeline	TX	W
St Michael's Academy	Austin	TX	W
St Philip School	El Campo	TX	W
Sugar Mill Elementary School	Sugar Land	TX	W
Trinidad School	Trinidad	TX	W
Trinity Christian School	Cedar Hill	TX	W
Ware Elementary Schooll	Longview	TX	W
Zavala Magnet School	Odessa	TX	W
Heritage JR/SR High School	Provo	UT	W
Hillfield Elementary School	Clearfield	UT	W
Mountain Shadows Elementary School	West Jordan	UT	W

School Name	City	State	Region
Riverview Junior High School	Murray	UT	W
Rosslyn Heights Elementary School	Salt Lake City	UT	W
San Juan High School	Blanding	UT	W
San Rafael Junior High School	Ferron	UT	W
Whitehorse High School	Montezuma Cree	UT	W
Brock Road Elementary School	Spotsylvania	VA	SE
Cave Spring Elementary School	Roanoke	VA	SE
Phoebus Senior High School	Hampton	VA	SE
Potomac High School	Dumfries	VA	SE
Sherwood Forest Elementary School	Norfolk	VA	SE
Thomas Harrison Middle School	Harrisonburg	VA	SE
Volens Elementary School	Nathalie	VA	SE
Charleston Elementary School	W Charleston	VT	NE
Brewster Elementary School	Brewster	WA	W
Coupeville Elementary School	Coupeville	WA	W
Coupeville High School	Coupeville	WA	W
Coupeville Middle School	Coupeville	WA	W
East Valley Central Middle School	Yakima	WA	W
Ebenezer Christian School	Lynden	WA	W
Manson Elementary School	Manson	WA	W
McCleary Elementary School	McCleary	WA	W
Monticello Middle School	Longview	WA	W
Odessa School	Odessa	WA	W
Steptoe Elementary School	Steptoe	WA	W
Toppenish Middle School	Toppenish	WA	W
Assumption High School	Wisconsin Rapids	WI	MW
Barton Elementary School	West Bend	WI	MW
Burroughs Middle School	Milwaukee	WI	MW
Immanuel Lutheran School	Greenville	WI	MW
James Madison Memorial High School	Madison	WI	MW
Malone Elementary School	Prescott	WI	MW

School Name	City	State	Region
Mattoon Elementary School	Mattoon	WI	MW
Muir Middle School	Milwaukee	WI	MW
Parker High School	Janesville	WI	MW
Queen of the Holy Rosary	Necedah	WI	MW
St Henry School	Watertown	WI	MW
St Joan of Arc School	Okauchee	WI	MW
St John the Baptist School	Seymour	WI	MW
Ashton Elementary School	Ashton	WV	SE
Kenna Elementary School	Charleston	WV	SE
Marlington Elementary School	Marlington	WV	SE
Arvada-Clearmont High School	Clearmont	WY	W

Index

A

C

D

E

M

N

O

P

R

S

T

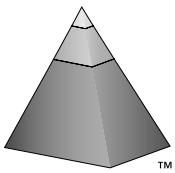
U

V

W

Y

Z



Renaissance Learning
P.O. Box 8036, Wisconsin Rapids, WI 54495-8036
(800) 656-6740
FAX: (715) 424-4242
Email: answers@renlearn.com
Web: www.renlearn.com