STAR Reading™

Understanding Reliability and Validity

Accurate Reading Assessment in Just 10 Minutes

STAR READING

One of the largest obstacles teachers face is the inability to quickly identify the strengths and weaknesses in pupils' reading performance. In other words, it is difficult to help pupils become better readers if you cannot determine their level of achievement, and when they need additional teaching.

STAR Reading™ spells the end of "hit-or-miss placement". This computer-adaptive test of reading comprehension provides accurate scores on demand. It works across the common academic years, years 2–13. With STAR Reading it is also possible to assess pupils' reading achievement frequently throughout the school year to track growth with no additional cost.

The STAR Reading assessment measures reading comprehension, and reports pupils' Reading Age (RA) and Estimated National Curriculum Level (NCL). It provides information to help teachers tailor teaching, monitor reading growth and improve pupils' reading performance. In approximately 10 minutes, this computer-adaptive test provides accurate reading scores for pupils in years 2–13.

Being able to identify your pupils' reading skills helps take the frustration out of improving reading skills. It allows you to guide your pupils to materials that they can accomplish without struggling, while still being challenged enough to strengthen their skills. It also helps you create teaching materials to present information at a level that your pupils are sure to understand. Knowing how your pupils compare to others helps you identify their own strengths and weaknesses, as well as any patterns of behaviour that you can use to develop stronger skills in deficient areas.

This brochure should make some difficult concepts easier to understand because it helps you:

- Find the correlation between STAR Reading and other national/standardised tests used in the UK and the US.
- Understand the reliability of the STAR Reading test, the standard error of measurement (SEM) and the validity of testing with STAR Reading.
- Learn how to interpret STAR Reading test scores.

STAR Reading is not intended for use as a national test. Rather, it is an assessment that can be used throughout the year to improve teaching, to increase learning and to better prepare for year-end tests while there is still time to improve performance before the regular testing cycle. For more information about STAR Reading, please call 0845 260 3570.

Test Content and Format

A STAR Reading test consists of 25 questions, or items. These questions are selected from a bank of more than 1,000 multiple-choice, vocabulary-in-context questions, appropriate for years 2–13.

The vocabulary-in-context items consist of a single sentence with a blank to indicate a missing word. The pupil must read and complete the sentence, choosing the correct word from a multiple-choice list of three or four words. Vocabulary-in-context items measure comprehension by requiring pupils to rely on background information, apply vocabulary knowledge and use active

strategies to construct meaning from the assessment text. These cognitive tasks are consistent with what researchers and practitioners describe as reading comprehension.

Test Reliability

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 STAR Reading 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. While developing national norms for STAR Reading 2.0 in the US, data were collected to allow all four of these methods to be applied. Reliability estimates were also developed in a separate study conducted within the UK, as a check on the US values.

The reliability data from the four US analyses are presented next. Following that, the reliability data from the UK are presented.

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 reliability. These split-half reliability coefficients are independent of the generic reliability approach discussed later in this section 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 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 for grades 1–12 displayed in the fourth column of Table 1.

As the data in this table indicate, sample sizes varied by grade, and ranged from 594 to 3,720 pupils. The reliability coefficients within grade ranged from .89 to .93.

Table 1: Scale Score Reliability Estimates

Grade	US Norming Sample			Test-Retest Sample		Alternate Forms Sample	
	Sample Size	Generic Reliability	Split-Half Reliability	Sample Size	Retest Reliability	Sample Size	Alternate Forms Reliability
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
Overall	29,169	0.96	0.96	2,095	0.94	4,551	0.95

Test-Retest Reliability Study

The test-retest study provided estimates of STAR Reading reliability using a variation of the test-retest method. In the traditional approach to test-retest reliability, pupils take the same test twice, with a short time interval, usually a few days, between administrations. In contrast, the STAR Reading 2.0 test-retest study administered two different tests by avoiding the use of any items on the second test that the pupil 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 1 and 2 makes the test-retest study a kind of alternate forms reliability study, but that term is reserved for another study, described later.)

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 2.0 test were calculated using the STAR Reading IRT (Item Response Theory) Rasch 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 2.0 test-retest reliability study consisted of a total of 2,095 pupils—a reasonable number of pupils for these kinds of analyses.

It is important to note that very little time elapsed between the first and second administrations of the pupils' 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 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 2.0 US norming software was constrained to start both tests at the same point. This helped maximise the parallelism of the two tests.

Reliability coefficients estimated from the test-retest study are provided in the sixth column of Table 1. 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 1 shows, the test-retest reliability estimated over all 12 grades was 0.94. Estimates by grade, which are smaller because score variances within grades are smaller, range from 0.79 to 0.91. Their average is 0.85—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 test. 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 provided an opportunity to develop estimates of STAR Reading alternate forms reliability. Pupils in this study took both a STAR Reading 2.0 US norming test and an original STAR Reading 1.x test, with an interval of days between tests. Order of administration was counterbalanced, with some pupils taking the STAR Reading 1.x test first, and the others taking the STAR Reading 2.0 norming 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 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 1–12 in the rightmost column of Table 1. As the data in this table indicate, the correlation was .95 for the overall sample of 4,551 pupils. 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 tests but also to their equivalence as measures of reading performance.

Generic Reliability Study

The data of the US norming study as a whole provided the opportunity to estimate what are referred to as generic reliability coefficients for the STAR Reading test. Estimates of generic reliability are derived from an IRT-based feature of the STAR Reading test: individual estimates of measurement error, called conditional SEMs, that are computed along with each pupil's Rasch IRT ability estimate, theta. Item Response Theory, and hence STAR Reading 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

$$1 - (\text{error variance}/\text{total score variance})$$

we can compute a reliability estimate by substituting the average of the individual pupil error variances (as the error variance term) and the variance of the pupils' 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 2.0 US norming data resulted in the generic reliability estimates shown in the third column of Table 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 adaptive test than are the more conservative retest and alternate forms coefficients.

The generic reliability estimates listed in Table 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 test.

Standard Error of Measurement

When interpreting the results of any test instrument, it is important to remember that the scores represent estimates of a pupil'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 pupil were tested repeatedly, his or her scores would fluctuate within 36 points of their first score about 68 per cent of the time, and within 72 points (twice the

SEM) roughly 95 per cent 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 test differs 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 pupil 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 pupil 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 pupil's ability level, the average SEMs for the IRT ability estimates are very similar for all pupils. However, because the transformation of the IRT ability estimates into equivalent Scale Scores is not linear, the SEMs in the Scale Score metric are less similar.

Table 2 summarises the average SEM values for the US 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 Scale Score SEMs. As the data indicate, the average IRT score SEMs are nearly constant regardless of grade level. In contrast, the SEMs of the Scale 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 Scale 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 2: Standard Errors of Measurement (IRT Scores and Scale Scores) STAR Reading Norming Analysis—Spring 1999

Grade	US Norming Sample Size	Average IRT Score SEM x 100	Average Scale Score SEM	5th Percentile Scale Score SEM	95th Percentile Scale Score SEM
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
Overall	29,169	49	74	21	137

Reliability Evidence from the UK

A large-scale validity study of STAR Reading was conducted in 28 schools in England in 2006. Pupils from years 2–9 were tested on STAR Reading to investigate the reliability of scores. Estimates of generic reliability were obtained from completed assessments. A random selection of pupils at each year was obtained to participate in a test-retest reliability analysis. In addition to the reliability estimates, the conditional standard error of measurement was computed for each individual pupil and summarised by year.

Results of the reliability analyses are found in Table 3. Generic reliability estimates range from a low of 0.92 in year 7 to a high of 0.96 in years 2 and 3. Test-retest reliabilities were obtained, on average, from 5–9 days after the initial testing. Results indicated high levels of score consistency over this time interval with test-retest reliabilities ranging from 0.69 in years 7 and 8 to 0.79 in year 4. The average conditional standard error of measurement increased with year from 26 scale score units in year 2 to 81 scale score units in year 9. Overall, these results indicated a high level of score consistency for a single assessment and on repeated occasions.

Table 3: Generic Reliability and Conditional SEM Estimates by Year in the UK

Year	Generic		Conditional SEM		Test-Retest		
	Sample Size	Estimate	Average	St. Dev.	Sample Size	Estimate	Avg. Days
2	557	0.96	26	20	49	0.75	5
3	1,076	0.96	34	17	125	0.75	9
4	1,439	0.94	46	19	157	0.79	6
5	1,514	0.94	54	25	136	0.71	7
6	1,229	0.93	64	27	126	0.76	6
7	4,029	0.92	72	29	350	0.69	9
8	1,480	0.93	77	32	151	0.69	9
9	632	0.93	81	31	61	0.70	6

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 comprehension, in the case of the STAR Reading test. These content validity issues were an integral part of the design and construction of the STAR Reading test items.

STAR Reading Version 4.1 UK employs vocabulary-in-context test items to measure reading comprehension. Each pupil's test is assembled adaptively from a bank of approximately 1,000 items whose difficulty has been calibrated using the Rasch IRT model. Items were reviewed by an expert team of content analysts in the US and by NFER in the UK to ensure all items were content relevant and culturally appropriate.

Each of those 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 year 1–2 reading level, three possible answers are given. Questions at a year 3 reading level and higher offer four possible answers.
2. To answer the question, the pupil selects the word that best completes the sentence from the answer choices. 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 minimise the amount of re-reading required.
4. The sentence provides sufficient context clues for pupils to determine the appropriate answer choice. However, the length of each sentence varies according to the guidelines shown in Table 4.
5. Typically, the words that provide the context clues in the sentence have a lower reading level than the actual test word. However, because the number of words that are available below year 3 are limited, not all of the questions at or below year 3 meet this criterion. Nonetheless, even at levels below year 3, no context words have a higher reading level than the reading level of the item itself.
6. The correct answer option is a word selected from the appropriate year level of the item set. Incorrect answer choices are words at the same test level or one year below. Through vocabulary-in-context test items, STAR Reading requires pupils to rely on background information, apply vocabulary knowledge and use active strategies to construct meaning from the assessment text. These cognitive tasks are consistent with what researchers and practitioners describe as reading comprehension.

Table 4: Maximum Sentence Length per Item Grade Level (Including Sentence Blank)

US Grade	Equivalent UK Year	Maximum Sentence Length
Kindergarten/Grade 1	Years 1 and 2	10 words
Grades 2 and 3	Years 3 and 4	12 words
Grades 4–6	Years 5–7	14 words
Grades 7–13	Years 8–14	16 words

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 test claims 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 year levels and improves as a pupil's year placement increases, STAR Reading scores should demonstrate these anticipated internal relationships; in fact, they do. Additionally, STAR Reading scores should correlate highly with other accepted procedures and measures that are used to determine reading achievement level; this is external validity.

Item Recalibration Results

To further evaluate the extent to which the items in STAR Reading are appropriate for UK pupils, an analysis of the item level data was undertaken. The analysis

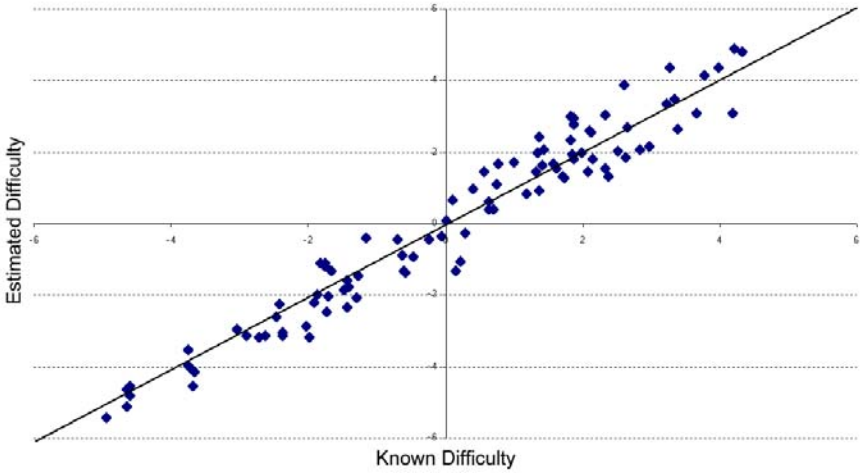
proceeded by recalibrating items from the entire STAR Reading database of UK users. A random selection of about 10 per cent of the short, vocabulary-in-context items were chosen, which resulted in 97 items for analysis. To recalibrate the items, pupil-ability estimates were used to anchor the scale, and item-difficulty estimates were obtained on those items. The recalibrated item difficulties in the UK sample were then compared to the known item difficulties of the STAR Reading items from the US calibration sample.

Results indicated that there was an average of 1,601 responses to each item with a standard deviation of 748. The minimum number of responses to an item was 500 and the maximum was 3,406. The average percentage correct across the items was 68 per cent, which was in-line with the expected value of 67.5 per cent. The average item-total correlation (point-biserial) was 0.37 with a standard deviation of 0.05.

The average difficulty of the known difficulty values was 0.11 with a standard deviation of 2.41, and the estimates in the UK sample had a mean of 0.03 and a standard deviation of 2.61. Using the distribution of the known values, the standardised mean differences between the estimates and known values was −0.03. This indicated a small practical difference in estimated difficulties.

The plot of the estimated values against the known values appeared linear (Figure 1). The correlation between known difficulties and the obtained estimates was 0.97. Regression analysis was used to evaluate the extent to which the estimated values differed in difficulty (intercept) and scaling (slope) features. The regression model was significant, $F(1, 95) = 1{,}544.29$, $p < 0.001$, $R^2 = 0.94$. The intercept was found to not significantly differ from zero, $t(1) = -1.64$, $p > 0.05$, and the slope was not found to differ significantly from unity, $F(1, 95) = 2.99$, $p > 0.05$. These results indicated a high level of linear correspondence between the UK estimates and the known values along with similarity in scaling and difficulty.

Figure 1: Scatterplot of the Difficulty Values Estimated in the UK Sample to the Known Difficulties



Concurrent and Predictive Validity Evidence from the UK

Evidence of substantial degrees of correlation between STAR Reading and scores on external tests has been gathered both in the UK and the US. In the UK, STAR Reading reports not only scale scores but also estimates of pupils' Reading Age and National Curriculum Levels. This section presents evidence of correlation of STAR Reading with widely-used UK tests; it also provides evidence of the accuracy of STAR Reading's estimates of RA and NCL. Sections following this one provide data from the US on the relationships of STAR Reading and dozens of other external tests.

Concurrent Validity Evidence

As STAR Reading is a vertically scaled assessment, it is expected that scores will increase over time and provide adequate separation between contiguous years to appropriately index developmental increases in reading comprehension. Descriptive statistics for test-result distributions of the UK pupils from the 28 schools that participated in the reliability study are found in Table 5. Results in Table 5 indicate that the median score (50^{th} percentile rank) and all other score distribution points increased systematically with year. In addition, a single-factor ANOVA was used to evaluate the statistical significance of mean differences in reading outcomes for each year. The results indicated significant differences between years, $F(7,11948) = 905.22$, $p < 0.001$, $\eta^2 = 0.35$, with observed power of 0.99. Follow-up analyses using Games-Howell post-hoc testing found significant differences, $p < 0.001$, between all years. These results provided confirmatory evidence of the developmental hypothesis, as significantly different reading outcomes were found across the different year cohorts.

Table 5: Descriptive Statistics for Pupil Test Performance in Scale Scores

Year	Sample Size	Percentile Rank				
		5	25	50	75	95
2	557	60	72	106	251	456
3	1,076	66	98	228	350	508
4	1,439	78	234	360	469	650
5	1,514	87	294	433	554	811
6	1,229	149	393	510	662	983
7	4,029	228	449	585	768	1,119
8	1,480	198	470	653	901	1,222
9	632	241	513	711	953	1,258

In addition, the time to complete a STAR Reading assessment was computed, to provide evidence for the length of time a test session lasted. The distribution of test times is provided in Table 6 by year and described by percentile rank. Results indicated at least half of the pupils at each year finished within 8 minutes while at least 75 percent finished within 10 minutes. Total test time also decreases with each subsequent year.

Table 6: Total Test Time, in Minutes, for a STAR Reading Test by Year, Given in Percentiles

Year	Sample Size	Percentile Rank				
		5	25	50	75	95
2	557	3.18	5.74	7.55	9.87	14.50
3	1,076	2.99	5.45	7.11	8.77	11.92
4	1,439	3.88	5.38	6.60	7.90	10.48
5	1,514	3.57	5.05	6.38	7.70	10.15
6	1,229	3.62	4.93	5.98	7.16	9.43
7	4,029	3.57	4.80	5.82	7.00	8.98
8	1,480	3.12	4.55	5.58	6.75	8.88
9	632	3.20	4.38	5.32	6.50	8.59

The National Foundation for Educational Research (NFER) conducted a large-scale validity study of STAR Reading in 28 schools in England in 2006.[1] Pupils in both primary (N=1,968) and secondary (N = 1,034) schools were recruited. The study investigated the concurrent validity of STAR Reading with a well-known and highly reliable test of reading comprehension that was developed and normed in the UK, the Suffolk Reading Scale 2 (SRS2).[2] Specific results of the study will be outlined below. The final results indicated a strong correlation (r = 0.91) between STAR Reading and SRS2, STAR Reading and Reading Ages (r = 0.91), and between STAR Reading and teacher assessments (TA) of pupil performance with respect to the English National Curriculum Levels (r = 0.85). The NFER report concluded that STAR Reading correlated highly with UK tests of reading "demonstrating concurrent evidence of their validity for use in this country" (pg. 21).

Specific results of the NFER study are found in Table 7. For the study, primary and secondary pupils took both STAR Reading and one of three age-appropriate forms of the Suffolk Reading Scale 2 in the fall of 2006. Pupils in years 2–3 were administered the Suffolk Scale form level 1A, pupils in years 4–6 were administered level 2A and pupils in years 7–9 were administered level 3A. Since pupils within non-overlapping year spans all took the same SRS2 form, the number correct score was used in the analyses. Pupil Reading Ages, given in months of age, were computed from their age and Suffolk reading scores. In addition to gathering external test data from the SRS2, teachers conducted individual assessments of each pupil's attainment in terms of the National Curriculum Levels in England, a measure of developmental progress that spans the primary and secondary years.

1. Sewell, J., Sainsbury, M., Pyle, K., Keogh, N., & Styles, B. (2007). *Renaissance Learning equating study report*. Slough, England: National Foundation for Education Research (NFER).
2. nferNelson (compiled by F. Hagley). (2002). *Suffolk reading scale 2*. London: nferNelson.

Table 7: Correlations of STAR Reading with Scores on the Suffolk Reading Scale 2 and Teacher Assessments in a Study of 28 Schools in England

School Years	Average Test Time	Suffolk Reading Scale 2				Teacher Assessments	
		Test Form	Sample Size	SRS2 Score[a]	Reading Age	Sample Size	Assessment Levels
2–3	7.3	SRS1A	713	0.84	0.85	n/a	n/a
4–6	6.0	SRS2A	1,255	0.88	0.90	n/a	n/a
7–9	5.4	SRS3A	926	0.78	0.78	n/a	n/a
Overall			2,894	0.91	0.91	2,324	0.85

a. Correlations with the individual SRS forms were calculated with within-form raw scores. The overall correlation was calculated with a vertical scale score.

Correlations between STAR Reading with all three measures are displayed in Table 7, by year band grouping and overall. Table 7 also provides the average time (in minutes) to complete STAR Reading for each year band. As the table indicates, the overall correlation between STAR Reading and SRS2 scores was 0.91, the correlation with Reading Age was 0.91, and the correlation with teacher assessments was 0.85. Within-form correlations with the SRS2 ability estimate ranged from 0.78 to 0.88, with a median correlation of 0.84, and ranged from 0.78 to 0.90 on Reading Age, with a median of 0.85.

The average time to complete the STAR Reading assessment was very similar to previous findings (see Table 6). The results indicated that most pupils were able to complete the assessment in a relatively short period of time. For instance, most primary school pupils were able to complete the assessment in less than 10 minutes.

Overall, these results provided evidence for the validity of STAR Reading scores for use in the UK. Evidence indicated that scores on STAR Reading were highly correlated with pupil outcomes with respect to their standing on the National Curriculum Levels. In addition, STAR Reading tests take a relatively short amount of time to administer. Therefore, STAR Reading provides a quick but powerful way to validly assess performance in the area of reading achievement.

Predictive Validity Evidence

Evidence of predictive validity was carried out in the 28 schools in England that were part of the reliability study. Pupils were initially tested on STAR Reading during October and November 2006. Follow-up assessments on STAR Reading were then completed on a subset of the pupils at the end of the academic year during the last 2 weeks of May, June and the first two weeks of July to obtain an end-of-year reading achievement outcome. The average number of months between testing occasions was between 7.6 months at years 2 and 3 and 8.3 months at year 7. Predictive validity coefficients ranged between 0.75 at year 8 and 0.86 at year 6. The overall correlation was 0.88, with an average of about 8 months between test occasions.

Table 8: Predictive Validity between Pretest and Posttest Results and the Average Months between Tests

Year	Sample Size	Correlation	Avg. Months
2	122	0.82	7.6
3	546	0.84	7.6
4	529	0.86	7.7
5	639	0.81	7.7
6	445	0.86	7.9
7	1,555	0.82	8.3
8	609	0.75	8.2
9	138	0.85	7.8
Overall	4,583	0.88	8.0

External Validity Evidence from the US

During the STAR Reading US 2.x norming study, schools submitted data on how their pupils performed on several other popular standardised test instruments along with their pupils' STAR Reading results. This data included more than 12,000 pupil 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 state tests.

Computing the correlation coefficients was a two-step process. First, where necessary, data were placed onto a common scale. If Scale 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). Scale Scores or NCE scores were then used to compute the Pearson product moment correlation coefficients.

Tables 9–12 present the correlation coefficients between the STAR Reading US 2.x test and each of the other test instruments for which data were received. Tables 9 and 10 display "concurrent validity" data, that is, correlations between STAR Reading 2.0 US norming study test scores and other tests administered at close to the same time. Tests listed in Tables 9 and 10 were administered during the spring of 1999, the same quarter in which the STAR Reading 2.0 US norming study took place. Tables 11 and 12 display 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 pupil records.

Tables 9 and 10 are connected and Tables 11 and 12 are connected, both sets being separated by US grade. Tables 9 and 11 present validity coefficients for US grades 1–6, and Tables 10 and 12 present the validity coefficients for US grades 7–12. The

bottom of each table presents a US grade-by-grade summary, including the total number of pupils for whom test data were available, the number of validity coefficients for that US 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 US grades 1–6, and .68 for US grades 7–12. The other validity coefficient within-grade averages varied from .60 to .77; the overall average was .73 for US grades 1–6, and .71 for US grades 7–12.

The extent that the STAR Reading US 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 standardised reading tests. The actual relationship between the STAR Reading test and the other tests is likely a bit higher than these estimates indicate.

Table 9: Concurrent Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Spring 1999, Grades 1–6.[a]

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

Table 9: Concurrent Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Spring 1999, Grades 1–6.[a] (Continued)

Test Form	Date	Score	1		2		3		4		5		6	
			n	r	n	r	n	r	n	r	n	r	n	r
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 (SAT-9)														
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*
TerraNova														
	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*
Summary														
Grade(s)		All	1		2		3		4		5		6	
Number of pupils		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												

a. Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level. Sample sizes are in the columns labelled "n".

Table 10: Concurrent Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Spring 1999, Grades 7–12.[a]

Test Form	Date	Score	7		8		9		10		11		12	
			n	r	n	r	n	r	n	r	n	r	n	r
California Achievement Test (CAT)														
/ 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 (SAT-9)														
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*	–	–	–	–
Wide Range Achievement Test 3 (WRAT3)														
	Spr 99		–	–	17	0.81*	–	–	–	–	–	–	–	–
Summary														
Grade(s)		All	7		8		9		10		11		12	
Number of pupils		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												

a. Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level. Sample sizes are in the columns labelled "n".

Table 11: Other External Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6.[a]

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

Table 11: Other External Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6.[a] (Continued)

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

Table 11: Other External Validity Data: STAR Reading 2.x Correlations (r) with External Tests Administered Prior to Spring 1999, Grades 1–6.[a] (Continued)

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

a. Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level. Sample sizes are in the columns labelled "n".

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

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

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

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

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

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

a. Asterisks (*) denote correlation coefficients that are statistically significant at the .05 level. Sample sizes are in the columns labelled "n".

Types of Test Scores

After pupils have tested with STAR Reading, the software uses their test results to determine specific test scores. When evaluated individually, these scores identify different characteristics of the pupils' reading performance. However, when you look at all of the scores as whole, you get an extremely accurate portrait of what the pupils are doing well—and where they need more concentrated attention.

Scale Score (SS)

STAR Reading software creates a virtually unlimited number of test forms as it dynamically interacts with the pupils 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 software does this in two steps. First, maximum likelihood is used to estimate each pupil'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 Scale Scores. STAR Reading Scaled Scores range from 0–1400.

Estimated Reading Age (Est. RA)

The Estimated Reading Age indicates the typical reading age for an individual with a given STAR Reading scale score. This provides an estimate of the chronological age at which pupils typically obtain that score. The Est. RA score is an approximation based on the demonstrated relationship between STAR Reading

and other tests of pupil reading ability, which were normed in the UK (see for instance the NFER report of concurrent validity).

The scale is expressed in the following form: YY:MM, where YY indicates the Reading Age in years and MM the months. For example, an individual who has obtained a Reading Age of 7:10 would be estimated to be reading as well as the average individual at 7 years, 10 months of age. Due to the range of items in STAR Reading and the intended range of years appropriate for use, a Reading Age cannot be determined with great accuracy if the reading ability of the pupil is either below 6:00 or above 16:06. Therefore, pupils who obtain an Est. RA of 6:00 should be considered to have a Reading Age of 6 years, 0 months or lower, and an Est. RA of 16:06 indicates a Reading Age of 16 years, 6 months or older.

Estimated National Curriculum Level–Reading (Est. NCL–Reading)

The Estimated National Curriculum Level in Reading is an estimate of a pupil's standing on the National Curriculum based on his or her STAR Reading performance. This score is an approximation based on the demonstrated relationship between STAR Reading scale scores and teachers' judgments, as expressed in their teacher assessments (TA) of pupils' attained skills. This score should not be taken to be the pupil's actual National Curriculum Level, but rather an estimate of the curriculum level at which the pupil is most likely performing. Stating this another way, the Est. NCL from STAR Reading is an estimate of the individual's standing in the national curriculum framework based on a modest number of STAR Reading test items, selected to match the pupil's estimated ability level. The estimated score is meant to provide information useful for decisions with respect to a pupil's present level of functioning; a pupil's actual NCL is obtained through national testing and assessment protocols.

The Est. NCL score is reported in the following format: the Estimated National Curriculum Level followed by a sublevel category, labelled a, b or c. The sublevels can be used to monitor pupil progress more finely, as they provide an indication of how far a pupil has progressed within a specific National Curriculum Level. For instance, a pupil with an Est. NCL of 4c would indicate that an individual is estimated to have just obtained level 4, while another pupil with a 4a is estimated to be approaching level 5.

It is sometimes difficult to identify whether or not a pupil is in the top of one level (for instance, 4a) or just beginning the next highest level (for instance, 5c.) Therefore, a transition category is used to indicate that a pupil is performing around the cusp of two adjacent levels. These transition categories are identified by a concatenation of the contiguous levels and sublevel categories. For instance, a pupil whose skills appear to range between levels 4 and 5, indicating he or she is probably starting to transition from one level to the next, would obtain an Est. NCL of 4a/5c. These transition scores are provided only at the junction of one level and the next highest. There are no transition categories within a level; for instance there are no 4c/4b or 4b/4a categories.

Zone of Proximal Development (ZPD)

The Zone of Proximal Development defines the readability range from which pupils 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 pupil's ZPD score. Specifically, it relates the Grade

Equivalent estimate of a pupil's reading ability with the range of most appropriate readability levels to use for reading practice.

The Zone of Proximal Development is especially useful for pupils 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 pupils in the 1996–1997 school year. More information is available in *ZPD Guidelines: Helping Students Achieve Optimum Reading Growth (2002), Teacher's Handbook 3–5: A Practical Guide to Reading Renaissance in the Intermediate Grades (2002),* and *Teacher's Handbook 6–8: A Practical Guide to Reading Renaissance in Middle School (2001)*; all three documents were published by Renaissance Learning.

APPENDIX

Progress Monitoring Assessments

Renaissance Learning, Inc. is the leading provider of the breakthrough technology of progress monitoring assessments—software that provides primary- and secondary-school teachers with objective, timely and accurate information to improve reading, writing and maths. Teachers have traditionally been burdened with old paper record-keeping systems. Now our family of progress monitoring assessments provides teachers with vastly improved information on pupil learning, freeing teachers to spend more quality time teaching. Progress monitoring assessments help teachers develop critical thinkers and lifelong learners—pupils who like maths and love to read. Research shows that pupils of teachers who use our progress monitoring assessments do better on performance-based and standardised tests; have higher scores in reading, writing and maths; and have better attendance.

Renaissance Learning, Inc.

Renaissance Learning is the world's leading provider of computer-based assessment technology for primary and secondary schools. Adopted by more than 70,000 North American schools, Renaissance Learning's tools provide daily formative assessment and periodic progress-monitoring technology to enhance core curriculum, support differentiated instruction, and personalize practice in reading, writing, and math.

Our products help educators make the practice component of their existing curriculum more effective by providing tools to personalize practice and easily manage the daily activities for pupils of all ability levels. As a result, teachers using Renaissance Learning products accelerate learning, achieve higher test scores on state and national tests, and get more satisfaction from teaching.

Copyright Notice

32 Harbour Exchange Square
London E14 9GE

Tel: 020 7184 4000
Fax: 020 7538 2625
Email: info@renlearn.co.uk
Website: www.renlearn.co.uk