



# Understanding Reliability and Validity

---

**Accurate  
Maths  
Assessment  
in Just  
15 Minutes**

# STAR MATHS

---

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

STAR Maths™ spells the end of “hit-or-miss placement”. The STAR Maths assessment measures pupils' attainment in maths, and estimates their National Curriculum Levels. It provides information to help teachers tailor teaching, monitor growth and improve pupils' maths performance. In less than 15 minutes, this computer-adaptive test provides accurate maths scores for pupils in years 2–13. Repeated testing throughout the academic year to monitor achievement growth is possible and can be done at no additional cost.

Being able to identify your pupils' mathematics skills helps take the frustration out of learning Maths. 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.

STAR Maths serves three primary purposes:

- It provides quick and accurate estimates of pupils' mathematics skills, allowing teachers to give pupils materials appropriate to their abilities.
- It measures growth in maths skills and helps demonstrate the effectiveness of a maths teaching or intervention programme.
- It helps predict how pupils will perform on national tests while there is still time to intervene.

STAR Maths' unique powers of repeatability and flexible administration provide specific advantages for everyone in the education process:

- For teachers, STAR Maths facilitates individualised teaching by identifying pupils' current mathematics levels.
- For head teachers, STAR Maths provides regular, accurate reports on performance at the class, year and building level, as well as year-to-year comparisons.
- For education authority administrators and assessment specialists, STAR Maths furnishes a wealth of reliable and timely data on mathematics growth at each school.

It also provides a valid basis for comparing data across schools, years and special pupil populations.

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

- Find the correlation between STAR Maths and other standardised tests.
- Understand the reliability of the STAR Maths test, the standard error of measurement (SEM) and the validity of testing with STAR Maths.
- Learn accurate definitions of specific test scores.

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

## Test Content and Format

A STAR Maths assessment consists of 24 multiple-choice questions. The first 16 questions assess numeration concepts and computation; the following 8 questions assess word problems, approximations, statistics, data analysis and probability, shapes and space, measures and algebra. Questions are selected from a bank of more than 1,700 multiple-choice items.

STAR Maths adapts the difficulty level of each test according to a pupil's responses. If the pupil answers a question incorrectly, the next question will be easier. If the pupil answers correctly, the next question will be more difficult. The average length of time required to take a STAR Maths test is 13–14 minutes. Note that the time required for completing the test increases with ability level. Pupils performing at the 90th percentile take, on average, 20 minutes to complete the test. Pupils working at the 10th percentile average 8 minutes.

## Reliability

### Test Reliability and Measurement Precision

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

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

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

Because STAR Maths 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. Three direct methods were used to estimate the reliability of the STAR Maths computer-adaptive test: generic reliability, the split-half method, and the alternate forms method. In the course of developing national norms for STAR Maths 2.0 in the US, data were collected to allow all three 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 three US analyses are presented immediately below. Following that, the reliability data from the UK are presented.

## Generic Reliability

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

$$reliability = 1 - \frac{\sigma_{error}^2}{\sigma_{total}^2}$$

where  $\sigma_{error}^2$  is the variance of the errors of measurement, and  $\sigma_{total}^2$  is the variance of the test scores. In STAR Maths, the variance of the test scores is easily calculated from Scale Score data. The variance of the errors of measurement may be estimated from the conditional SEM statistics that accompany each of the IRT-based test scores, including the Scale Scores, as follows:

$$\sigma_{error}^2 = \frac{1}{n} \sum_n SEM_i^2$$

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

Using this technique with the data collected in the course of norming STAR Maths 2.0 in the US resulted in the generic reliability estimates shown in the third column of Table 1. Because this method is not susceptible to error variance introduced by repeated testing, multiple occasions and alternate forms, the resulting estimates of reliability are generally higher than the more conservative alternate forms reliability coefficients. These generic reliability coefficients are, therefore, plausible upper-bound estimates of the actual reliability of the STAR Maths computer-adaptive test.

**Table 1: Scale Score Reliability Estimates**

Grade	US Norming Sample			Alternate Forms Sample	
	Sample Size	Generic Reliability	Split-Half Reliability	Sample Size	Alternate Forms Reliability
1	3,076	0.83	0.82	745	0.73
2	3,193	0.79	0.78	866	0.75
3	2,972	0.80	0.78	853	0.74
4	2,981	0.81	0.79	840	0.73
5	3,266	0.83	0.80	813	0.79
6	2,555	0.84	0.84	729	0.73
7	2,896	0.86	0.86	698	0.72
8	2,598	0.88	0.88	714	0.74
9	1,771	0.86	0.86	381	0.79
10	1,556	0.88	0.87	304	0.80
11	1,419	0.87	0.87	255	0.76
12	945	0.88	0.88	191	0.72
Overall	29,228	0.95	0.94	7,389	0.91

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

## Split-Half Reliability

In classical test theory, before the advent of digital computers automated the calculation of internal consistency reliability measures such as Cronbach's alpha, approximations such as the split-half method were sometimes used. A split-half reliability coefficient is calculated in three steps. First, the test is divided into two halves, and scores are calculated for each half. Second, the correlation between the two resulting sets of scores is calculated; this correlation is an estimate of the reliability of a half-length test. Third, the resulting reliability value is adjusted, using the Spearman-Brown formula, 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 Maths reliability. These split-half reliability coefficients are independent of the generic reliability approach discussed earlier and more firmly grounded in the item-response data. Column 4 of Table 1 contains split-half reliability estimates for STAR Maths, calculated from the US norming study data.

## Alternate Forms Reliability Study

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

The alternate forms approach was used for STAR Maths, and the results are presented in the rightmost column of Table 1. Participating US schools were asked to administer two norming tests, each on a different day, to about one-fourth of the overall sample. Figure 1 is a scatter plot of their scores. This resulted in an alternate forms reliability subsample of more than 7,000 pupils who took different forms of the 24-item STAR Maths 2.0 US norming test. The interval between the first and second tests averaged four days. The interval varied widely, however. For example, in some cases both tests were given on the same day; in other cases, the interval ranged from one to as many as 40 days. Errors of measurement due to both content sampling and temporal changes in individuals' performance can affect alternate forms reliability coefficients, usually making them appreciably lower than internal consistency reliability coefficients. In addition, any growth in the trait that takes place in the interval between tests can also lower the correlation. The actual reliability of STAR Maths is probably higher than the alternate forms estimates presented in Table 1.

**Figure 1: Scatter Plot of Test Scores from the STAR Maths 2.0 US Norming Alternate Forms Reliability Study**

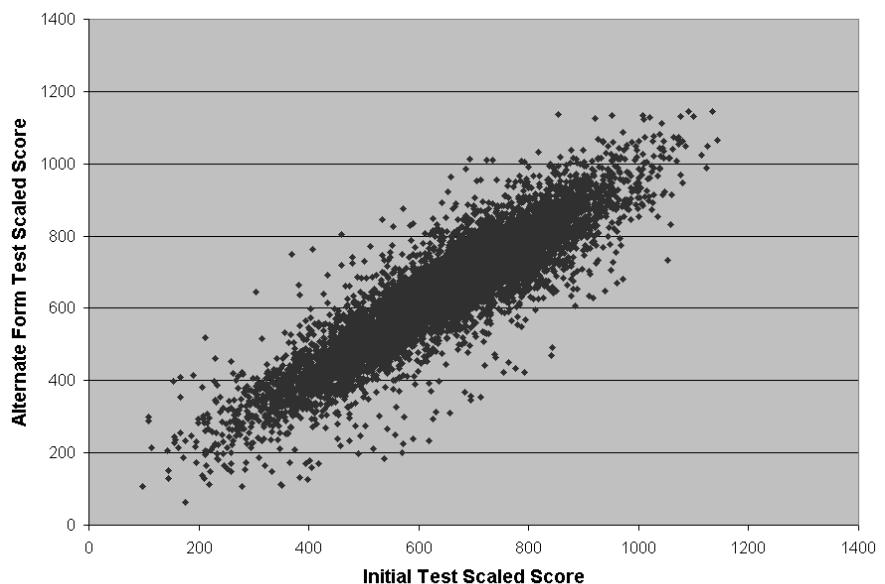


Table 1 lists the detailed results of the generic, split-half and alternate forms reliability analyses of STAR Maths 2.0 Scale Scores, both overall and by grade. The split-half and generic reliability estimates, which are based on the entire US norms sample of 29,228 pupils,<sup>1</sup> are very similar to one another, with the split-half values generally slightly lower. The aggregate of these reliability estimates is approximately 0.94. By grade, they range from 0.77 to 0.88, with a median of 0.85.

The alternate forms reliability estimates are based on the 7,389 pupils who participated in the reliability study, about one-fourth of the US norms sample. The aggregate alternate forms reliability was approximately 0.91. By grade, the values ranged from approximately 0.72 to 0.80, with a median value of 0.74.

## Standard Error of Measurement

When interpreting any educational test scores, the test user must bear in mind that the scores include some degree of error. The size of the test score reliability coefficient provides an indication of the overall magnitude of that error. The SEM arguably provides a measure that is more useful for score interpretation, as the SEM is expressed in the same units used to express the test score.

For the STAR Maths Scale Score, a conditional SEM is calculated for each individual, and the value of the SEM is included in the score reports, either explicitly or graphically. In the following section, aggregate SEMs are presented. For the Scale Score, these SEMs represent averages, overall and by year. Because the conditional SEMs vary systematically by Scale Score, the individual SEMs in the STAR Maths score reports are more useful for interpretation.

### Scale Score SEMs

The STAR Maths software calculates the SEM for each individual. This statistic is called the “conditional SEM” as it is conditional on the value of the Scale Score. Conditional SEMs vary from one pupil to another, and the interpretation of individual scores should be based on the pupil’s own SEM value. However, for purposes of summarizing the measurement precision of STAR Maths, average conditional SEM values are presented below. As the SEM estimates may vary with ability level, these SEM estimates will be tallied separately for each year, as well as overall.

Table 2 contains means and standard deviations (S.D.) of the STAR Maths 2.0 Scale Score conditional SEMs, overall and by grade, for the STAR Maths 2.0 US norms sample. The aggregate mean SEM value was 40, averaged over all grades. Within-grade averages range from 37 at grade 1 to 42 at grade 12.

---

1. There were 29,228 cases in the US norms sample; 43 with outlier scores were not included in the norms calculations, but were included in the reliability calculations.

**Table 2: STAR Maths 2.0 Standard Error of Measurement of Scale Scores**

Grade	Sample Size	Conditional SEM	
		Mean	S.D.
1	3,076	37	5.1
2	3,193	40	4.6
3	2,972	39	3.8
4	2,981	39	3.9
5	3,266	41	4.5
6	2,555	41	4.9
7	2,896	41	5.1
8	2,598	41	5.5
9	1,771	41	5.6
10	1,556	42	6.4
11	1,419	42	6.0
12	945	42	6.6
Overall	29,228	40	5.2

## Reliability Evidence from the UK

The National Foundation for Educational Research (NFER) conducted a large-scale validity study of STAR Maths in 28 schools in England in 2006.<sup>2</sup> Pupils from years 2–9 (N = 3,631) were tested on STAR Maths to investigate the reliability of scores. Estimates of generic reliability were obtained from completed assessments. A small subset of these pupils was selected to estimate test-retest reliability across all years. 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. The overall generic reliability across years was 0.94. Estimates within years ranged from a low of 0.87 in year 4, 5 and 9 to a high of 0.90 in years 2, 6, 7 and 8. Test-retest reliabilities were obtained on a small subset of pupils (N = 159) across the years who were retested, on average, about 4 days after the initial testing. Results indicated high levels of score consistency over this time interval with a test-retest reliability estimate of 0.74. The average conditional standard error of measurement was

2. Sewell, J., Sainsbury, M., Pyle, K., Keogh, N., & Styles, B. (2007). *Renaissance Learning equating study report*. Slough, England: National Foundation for Education Research (NFER).

stable across years and was on average about 36 scale score points with a standard deviation of 2 scale score points. 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	326	0.90	37	3			
3	351	0.89	36	4			
4	588	0.87	36	3			
5	467	0.87	36	2			
6	412	0.90	36	2			
7	680	0.90	36	3			
8	527	0.90	36	3			
9	280	0.87	36	2			
Overall	3,631	0.94	36	3	159	0.74	4

# VALIDITY

---

The key concept 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 is a difficult process because there are actually many aspects of validity that can be examined. For example, the content validity of the test deals with the relevance of the questions, strands and objectives sampled by the test.

Establishing construct validity involves using data and other information external to the test instrument itself. For example, the STAR Maths test claims to provide an estimate of a child's Mathematical achievement level for use in placement. Therefore, demonstration of STAR Maths' construct validity rests on the evidence that the test in fact provides such an estimate.

There are a number of ways to demonstrate this. One method includes examining the relationship between pupils' STAR Maths Scale Scores and their year levels. Since Mathematical ability varies significantly within and across year levels and improves as a pupil's year level increases, STAR Maths data should demonstrate these anticipated relationships.

Another source of evidence for construct validity is the relationship between pupils' STAR Maths scores and their scores on other measures of Mathematics achievement. If it is a valid assessment, the STAR Maths test should correlate highly with other accepted procedures and measures that are used to determine Mathematics achievement level. Among other things, pupils' STAR Maths scores should correlate highly with their scores on other established tests of Mathematics proficiency and achievement. Additionally, these scores should be highly related to teachers' assessments of their pupils' proficiency in Mathematics.

In the remainder of this section, information about the content of STAR Maths is presented, as evidence of its validity for assessing maths in UK schools. Following that, validity evidence that demonstrates a strong and positive correlation between STAR Maths scores and scores on other standardised tests will be presented.

## STAR Maths Content and Objective Clusters

STAR Maths test content is intended to reflect the objectives commonly taught in the Mathematics curricula of contemporary schools. The following major sources helped to define this curriculum content:

- National Curriculum (UK)
- National Numeracy Strategies (UK)
- National Foundation for Educational Research-NFER (UK organisation)
- Trends in International Mathematics and Science Study (TIMSS)
- Principles and Standards for School Mathematics of the National Council of Teachers of Mathematics (US organisation)
- Content specifications for the National Assessment of Educational Progress (US assessment)
- An extensive review of content covered in leading textbook series
- Curriculum guides and lists of objectives

There is reasonable, although not universal, agreement among these sources about the content of Mathematics curricula.

The final STAR Maths content specifications were intended to cover the objectives most frequently found in these sources. The STAR Maths content is organised into eight strands. There are 193 objectives within the eight strands.

## Numeration Concepts

The Numeration Concepts strand encompasses 42 objectives, making it the strand with the largest number of objectives. This strand concentrates on conceptual development of the decimal number system. At the lowest levels, it covers cardinal and ordinal numbers through ten (the ones). The strand then proceeds to treatment of the decades (tens), hundreds, thousands and then larger numbers such as hundred thousands and millions, all in the whole-number realm. At each of these levels of the number system, specific objectives relate to place value identification, number-numeral correspondence and expanded notation. Following treatment of the whole numbers, the Numeration Concepts strand moves to fractions and decimals. Coverage includes representation of fractions and decimals on the number line, conversions between fractions with different denominators and between fractions and decimals, number-numeral correspondence for decimals and rounding decimals.

Included in this category are specific objectives on roots, index notation, primes, composites and scientific notation. Because items in the Numeration Concepts strand emphasise understanding basic concepts, they are deliberately written to minimise computational burden.

## Computation Processes

The Computation Processes strand includes 39 specific objectives, the second largest number among the STAR Maths strands. This strand covers the four basic operations (addition, subtraction, multiplication and division) with whole numbers, fractions, decimals and percentages. Ratios and proportions are also included in this strand. Coverage of computational skill begins with the basic facts of addition and subtraction, starting with the fact families having sums to 10, then with sums to 18. The strand progresses to addition and subtraction of two-digit and three-digit numbers without regrouping, then with regrouping. At about the same level, basic facts of multiplication and division are introduced. Then, the four operations are applied to more difficult regrouping problems with whole numbers. Fractions are first introduced by way of addition and subtraction of fractions with like denominators. These are relatively easy for pupils in the US. However, the strand next includes operations with fractions with unlike denominators, mixed numbers and decimal problems requiring place change, all of which are relatively difficult for pupils. The Computation Processes strand concludes with a series of objectives requiring operations with percentages, ratios and proportions.

Although the Computation Processes strand can be subdivided into nearly an infinite number of objectives, the STAR Maths item bank provides a representative sampling of computational problems that cover the major types of problems pupils are likely to encounter. Indeed, the item bank does not purport to cover every conceivable computational nuance. In addition, among the more difficult problems involving computation with whole numbers, there are number combinations for which one would ordinarily use a calculator. However, it is

expected that pupils will know how to perform these operations by hand, and hence, a number of such items are included in the STAR Maths item bank.

The Numeration Concepts and Computation Processes strands are considered by many to be the heart of the basic Mathematics curriculum. Pupils must know the four operations with whole numbers, fractions, decimals and percentages. Pupils must know numeration concepts to have an understanding of how the operations work, particularly for regrouping, changing denominators in fractions and changing places with decimals and percentages. As noted above, these two strands constitute the first two thirds of the STAR Maths test. Mathematical development within these two strands also serves as the principal basis for teaching and learning recommendations provided in the STAR Maths Diagnostic Report. The remaining strands comprise the latter third of the STAR Maths test. This part might be labelled “applications” since many—although not all—of the objectives in this part can be considered practical applications of mathematical content and procedures. It is important to note that research conducted at the item-calibration stage of STAR Maths development demonstrated that the items in the various strands were strongly unidimensional, thus justifying the use of a single score for purposes of reporting.

## Approximations

The Approximations strand is also designed to parallel the Computation Processes strand in terms of the types of operations required. Again, many, but not all computation objectives are reflected in this strand. Obviously, in the Approximations strand, pupils are not required to compute a final answer. With number combinations similar to those represented in the Computation Processes strand, pupils are asked to approximate an answer. To discourage pupils from actually computing answers, response options are generally given in round numbers. The range of numerical values used in the options is generally set so that a reasonable approximate is adequate.

## Shape and Space

Although many curricular sources combine shape and space and measures in a single strand, the STAR Maths test represents them separately. At the lowest level, the Measures strand includes objectives on temperature and time (clocks, days of the week and months of the year). The strand provides coverage of both metric and customary (imperial) units. Metric objectives include use of the metric prefixes (milli-, centi-, etc.) and the conversion of metric and imperial units. The Measures strand also includes an objective on the measure of angles, one of the best examples of the overlap between the shape and space and measures areas.

## Measures

Although many curricular sources combine shape and space and measures in a single strand, the STAR Maths test represents them separately. At the lowest level, the Measures strand includes objectives on temperature and time (clocks, days of the week and months of the year). The strand provides coverage of both metric and customary (imperial) units. Metric objectives include use of the metric prefixes (milli-, centi-, etc.) and the conversion of metric and imperial units. The Measures strand also includes an objective on measures of angles, one of the best examples of the overlap between the shape and space and measures areas.

## Data Analysis and Probability

This strand begins with simple, straightforward extraction of information from tables, bar graphs and pie charts. In these early objectives, information needed to answer the question is given directly in the table, chart or graph. At the next higher level of complexity, pupils must combine or compare two or more pieces of information in the table, chart or graph in order to answer the question. This strand also includes several objectives related to probability and statistics. Curricular placement of probability and statistics objectives varies considerably from one source to another. In contrast, using tables, charts and graphs is commonly encountered across a wide range of years in nearly all Mathematics curricular materials.

## Word Problems

The Word Problems strand includes simple, situational applications of computations. In fact, the Word Problems strand is deliberately structured to parallel the Computation Processes strand in terms of the types of operations required.

Most computation objectives are paralleled in the Word Problems strand. For all items in the Word Problems strand, pupils are presented with a practical problem, and to answer the item correctly, they must determine what type of computational process to use and then correctly apply that process. The reading level of the problems is kept at a low level to ensure valid assessment of ability to solve word problems.

## Algebra

The final strand in the curricular structure of the STAR Maths item bank is Algebra. Although algebra is sometimes thought of as a higher level course, elements of algebra are actually introduced much earlier in the contemporary Mathematics curriculum. The use of simple number sentences and the translation of word problems into equations (at a very simple level) are introduced even in the lower years. Such objectives are included at the lowest level of the STAR Maths Algebra strand. The objectives progress rapidly in difficulty to those found in the formal algebra course. These more difficult objectives include operating with polynomials, quadratic equations and graphs of linear and non-linear functions.

## Objective Clusters

The STAR Maths Diagnostic Report contains two bar charts that reflect each pupil's performance on the Numeration Concepts and Computation Processes strands. By viewing these two charts, teachers can graphically see how each pupil is progressing in these two important areas. The report highlights these two strands because they form the foundation for the Mathematics curriculum. According to the National Council of Teachers of Mathematics' Principles and Standards for School Mathematics (NCTM), "understanding numbers and operations, developing number sense and gaining fluency in arithmetic computation form the core of Mathematics education for the US elementary grades" (p. 32).

The content in the Numeration Concepts and Computation Processes strands is organised in a hierarchical structure, reflecting the fact that pupils' mathematical development (and maths curriculum) proceeds in a step-like fashion. In other

words, their understanding of harder concepts is dependent upon their understanding the more basic concepts. For example, a pupil must first learn how to add numbers together before he or she is able to multiply them.

Because of this hierarchical structure and because every objective within these two strands could not be included on the STAR Maths 3.x and higher Diagnostic Report, for data reduction purposes, common objectives were grouped together, forming “objective clusters”. Based on the recommendations of a Mathematics content expert, the 42 Numeration Concepts objectives and the 39 Computation Processes objectives in STAR Maths 2.x and higher were grouped into 9 Computation and 8 Numeration clusters. The objectives included in each cluster in each strand are shown in Table 4.

**Table 4: Content of Objective Clusters for the STAR Maths Numeration Concepts and Computation Processes Strands**

Strand	Objective Cluster	Objective ID	Objective Name
Numeration Concepts	Ones	N00	Ones: Locate numbers on a number line
		NA1	Ones: Placing numerals in order
		NA2	Ones: Using numerals to indicate quantity
		NA3	Ones: Relate numerals and number words
		NA4	Ones: Use ordinal numbers
	Tens	N01	Tens: Place numerals (10–99) in order of value
		N02	Tens: Associate numeral with group of objects
		N03	Tens: Relate numeral and number word
		N04	Tens: Identify one more/one less across decades
		N05	Tens: Understand the concept of zero
	Hundreds	N06	Hundreds: Place numerals in order of value
		N07	Hundreds: Relate numeral and number word
		N08	Hundreds: Identify place value of digits
		N09	Hundreds: Write numerals in expanded form
Thousands	N11	Thousands: Place numerals in order of value	
	N12	Thousands: Relate numeral and number word	
	N13	Thousands: Identify place value of digits	
	N14	Thousands: Write numerals in expanded form	

**Table 4: Content of Objective Clusters for the STAR Maths Numeration Concepts and Computation Processes Strands (Continued)**

Strand	Objective Cluster	Objective ID	Objective Name
Numeration Concepts (continued)	Hundred Thousands	N16	Ten thousands, hundred thousands, millions, billions: Place numerals in order of value
		N17	Ten thousands, hundred thousands, millions, billions: Relate numeral and number word
		N18	Ten thousands, hundred thousands, millions, billions: Identify place value of digits
		N19	Ten thousands, hundred thousands, millions, billions: Write numerals in expanded form
	Fractions and Decimals	N21	Fractions and decimals: Convert fraction to equivalent fraction
		N22	Fractions and decimals: Convert fraction to decimal
		N23	Fractions and decimals: Convert decimal to fraction
		N24	Fractions and decimals: Read word names for decimals to thousandths
		N25	Fractions and decimals: Identify place value of digits in decimals
		N26	Fractions and decimals: Identify position of decimals on number line
		N27	Fractions and decimals: Identify position of fractions on number line
		N28	Fractions and decimals: Convert improper fraction to mixed number
		N29	Fractions and decimals: Round decimals to tenths, hundredths
		N30	Fractions and decimals: Relate decimals to percents
		Advanced Concepts I	N31
	N34		Advanced concepts: Recognise meaning of exponents (2–10)
	N39		Advanced concepts: Can determine greatest common factor
	N41		Advanced concepts: Recognises use of negative numbers

**Table 4: Content of Objective Clusters for the STAR Maths Numeration Concepts and Computation Processes Strands (Continued)**

Strand	Objective Cluster	Objective ID	Objective Name
Numeration Concepts (continued)	Advanced Concepts II	N32	Advanced concepts: Give approximate square roots of a number
		N33	Advanced concepts: Recognise the meaning of $n^{\text{th}}$ root
		N35	Advanced concepts: Recognise meaning of negative exponents
		N36	Advanced concepts: Recognise meaning of fractional exponents
		N37	Advanced concepts: Can use scientific notation
		N38	Advanced concepts: Knows meaning of primes and composites
		N40	Advanced concepts: Can determine least common multiple
Computation Processes	Addition and Subtraction Basic Facts to 10	C01	Addition of basic facts to 10
		C02	Subtraction of basic facts to 10
	Addition and Subtraction Basic Facts to 18, No Regrouping	C03	Addition of basic facts to 18
		C04	Subtraction of basic facts to 18
		C05	Addition of three single-digit addends
		C06	Add beyond basic facts, no regrouping ( $2d + 1d$ )
		C07	Subtract beyond basic facts, no regrouping ( $2d - 1d$ )
	Addition and Subtraction with Regrouping	C08	Add beyond basic facts with regrouping ( $2d + 1d, 2d + 2d$ )
		C09	Subtract beyond basic facts with regrouping ( $2d - 1d, 2d - 2d$ )
		C10	Add beyond basic facts with double regrouping ( $3d + 2d, 3d + 3d$ )
		C11	Subtract beyond basic facts with double regrouping ( $3d - 2d, 3d - 3d$ )
	Multiplication and Division: Basic Facts	C12	Multiplication basic facts
		C13	Division basic facts
		C14	Multiplication beyond basic facts, no regrouping ( $2d \times 1d$ )

**Table 4: Content of Objective Clusters for the STAR Maths Numeration Concepts and Computation Processes Strands (Continued)**

Strand	Objective Cluster	Objective ID	Objective Name
Computation Processes (continued)	Advanced Computation with Whole Numbers	C15	Division beyond basic facts, no remainders ( $2d \div 1d$ )
		C16	Multiplication with regrouping ( $2d \times 1d$ , $2d \times 2d$ )
		C17	Division with remainders ( $2d \div 1d$ , $3d \div 1d$ )
		C18	Add whole numbers: any difficulty
		C19	Subtract whole numbers: any difficulty
		C21	Divide whole numbers: any difficulty
	Fractions and Decimals I	C22	Add fractions: like single digit denominators
		C23	Subtract fractions: like single digit denominators
		C33	Add decimals, place change ( $2 + .45$ )
		C35	Subtract decimals, place change ( $5 - .4$ )
	Fractions and Decimals II	C24	Add fractions: unlike single digit denominators
		C25	Subtract fractions: unlike single digit denominators
		C26	Multiply fractions: single digit denominators
		C27	Divide fractions: single digit denominators
		C28	Add mixed numbers
		C29	Subtract mixed numbers
		C36	Multiply decimals
		C37	Divide decimals
	Percentages, Ratios and Proportions	C38	Percentage A (10 is what % of 40)
		C39	Percentage B (20% of 50 is what)
		C40	Percentage C (30 is 50% of what)
		C41	Proportions
		C42	Ratios
	Multiplication and Division of Mixed Numbers	C30	Multiply mixed numbers
		C31	Divide mixed numbers

On the STAR Maths Diagnostic Report, the shaded region of each bar chart reflects the amount of material within each strand that the pupil has most likely mastered. These estimates are based on the STAR Maths 2.0 norming data, and mastery is defined as 70 per cent proficient. Therefore, if a pupil's ability estimate suggests that he or she could answer 70 per cent or more correct on a specific objective cluster, such as Hundreds, he or she will have "mastered" that objective

cluster and that box will be shaded on his or her Diagnostic Report. Because the content in the strands included in the objective clusters is hierarchical, pupils most likely master the objective clusters in sequential order. The solid black line on the bar chart points to the objective cluster that the pupil is currently developing or the lowest objective that he or she has not mastered.

## Validity Evidence from the UK

In the NFER 2006 study, pupils in both primary (N = 2,006) and secondary (N = 883) schools were recruited. The study investigated the concurrent validity of STAR Maths with a well-known and highly reliable test of Mathematics ability that was developed and normed in the UK, Progress in Maths 4-14 Series published by nferNelson.<sup>3</sup> In addition, all participants received teacher assessments (TA) of their present Mathematics skills with respect to the National Curriculum Level in England. Specific results of the study will be outlined in this section.

As STAR Maths 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 mathematics achievement. Descriptive statistics for scale score distributions among the UK pupils that participated in the reliability study are found in Table 5.

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

Year	Sample Size	Percentile Rank				
		5	25	50	75	95
2	326	176	284	348	437	529
3	310	213	367	416	496	586
4	588	335	452	514	578	674
5	467	395	503	566	635	736
6	410	448	545	618	696	803
7	680	459	577	659	745	820
8	527	514	626	693	780	876
9	280	545	635	716	786	854

The correlation between STAR Maths scale score and pupil age at time of testing was 0.71. Results in Table 5 indicate that the median score (50th percentile rank) and all other score distribution points, except at the 95th percentile rank between year 8 and 9, gradually increase across years. In addition, a single-factor ANOVA was computed to evaluate the significance of differences between means at each

3. Clausen-May, T., Vappula, H., & Ruddock, G. (2004). *Progress in maths 4-14 series*. London: nferNelson.

year. The results indicated significant differences between years,  $F(7,3580) = 510.90$ ,  $p < 0.001$ ,  $\eta^2 = 0.50$ , with observed power of 0.99. Follow-up analyses using Games-Howell post-hoc testing found significant differences,  $p < 0.01$ , between all years, except years 8 and 9. These results provided confirmatory evidence of the developmental hypothesis, as Mathematics outcomes generally rise across years.

In addition, the time to complete a STAR Maths assessment was computed, to provide evidence of typical test duration. 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 11 minutes. Total test time also decreases with each subsequent year.

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

Year	Time to Complete a STAR Maths Test					
	Percentile Rank					
	Sample Size	5	25	50	75	95
2	326	4.46	8.05	10.78	15.61	25.78
3	351	4.83	7.79	10.50	13.70	21.11
4	588	5.69	8.47	11.03	14.37	20.64
5	467	5.48	7.90	10.72	14.30	20.13
6	412	5.67	7.87	10.08	13.45	19.00
7	680	5.00	8.07	10.57	13.68	20.22
8	527	5.00	7.57	9.53	11.93	17.62
9	280	4.35	6.98	8.78	11.10	16.68

## Concurrent Validity Evidence

For years 2–9 combined, the study found an overall correlation of 0.85 between STAR Maths and Progress in mathematics scores, and a correlation of 0.81 between STAR Maths and the teacher assessments (see Table 7). These are large correlations and provide strong evidence of the concurrent validity of STAR Maths scores.

All within-year correlations were between 0.73 and 0.75, except the one for year 2, with a median correlation of 0.74. The report authors concluded that STAR Maths correlated well with UK tests of maths, “demonstrating concurrent evidence of their validity for use in this country” (pg. 211).<sup>4</sup>

4. Sewell, J., Sainsbury, M., Pyle, K., Keogh, N., & Styles, B. (2007). *Renaissance Learning equating study report*. Slough, England: National Foundation for Education Research (NFER).

Average testing times were similar to the results from the reliability study, with average test times between 10–13 minutes.

**Table 7: Correlations of STAR Maths with Scores on the Progress in Mathematics Tests and Teacher Assessments in a Study of 28 Schools in England**

School Year	Progress in Mathematics				Teacher Assessments in Mathematics	
	Average Test Time	Test Form	Sample Size	PIM Score <sup>a</sup>	Sample Size	Assessment Levels
2	12.5	PiM 6	290	0.58		
3	11.5	PiM 7	297	0.73		
4	12.0	PiM 8	403	0.74		
5	11.4	PiM 9	415	0.74		
6	10.7	PiM 10	354	0.75		
7	11.0	PiM 11	257	0.74		
8	10.5	PiM 12	311	0.75		
9	9.1	PiM 13	233	0.73		
Overall			2,560	0.85	2,460	0.81

a. Correlations within PIM forms were calculated with age-standardised scores. The overall correlation was calculated with a vertically Scale Score developed by Renaissance Learning.

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

## Item Recalibration Results

To further evaluate the extent to which the items in STAR Maths were appropriate for UK pupils, an analysis of item-level data was undertaken. The analysis proceeded by recalibrating items from the entire STAR Maths database of UK users. A random selection of about 10 per cent of the items was chosen conditional on having over 300 valid responses, which resulted in 83 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 Maths items from the US calibration sample.

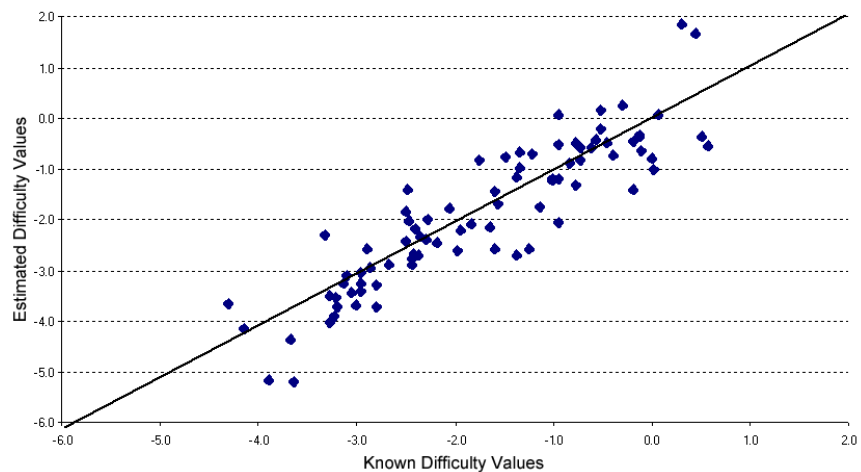
Results indicated the average number of responses per item was 464 with a standard deviation of 249. The minimum number was 303 with a maximum of

2,001 responses. The average percentage correct across the items was 66 per cent with a standard deviation of 14 per cent, which was consistent with the expectation of 67.5 per cent. The average item-total (point biserial) correlation was found to be 0.41 with a standard deviation of 0.06 with a minimum of 0.27 and a maximum of 0.58.

A scatter plot of the item difficulties estimated from the UK pupils against the known difficulty values can be seen in Figure 2. The overall correlation between estimated values and the known values was 0.90. The average difficulty of the known items was  $-1.73$  with a standard deviation of 1.23. The average difficulty of the estimated values was  $-1.88$  with a standard deviation of 1.42. The standardised mean difference between the estimated values and known values, using the known values distribution as the referent, was found to be of a small practical magnitude,  $d = -0.12$ .

Regression analysis was used to evaluate the extent to which the estimated values differed in difficulty (intercept) and scaling (slope) features. Regression of the estimated values on the known values was significant,  $F(1, 81) = 337.28$ ,  $p < 0.001$ ,  $R^2 = 0.81$ . Parameter estimates suggested the intercept was not significantly different from zero,  $t(1) = -0.61$ ,  $p > 0.10$ , and the slope was not significantly different from unity,  $F(1, 81) = 0.50$ ,  $p > 0.10$ . 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 2: Scatter Plot of the Item Difficulty Values Estimated in the UK Sample and the Known Difficulty Values**



## Predictive Validity Evidence

Evidence of predictive validity was collected in the 28 schools in England that were part of the reliability study. Pupils were initially tested on STAR Maths during October and November 2006. Follow-up assessments on STAR Maths were then completed on a subset of the pupils at the end of the academic year during the last 2 weeks of April and all of May to obtain an end-of-year mathematics achievement outcome.

Results presented in Table 8 indicated that all correlations between the beginning-of-the-year and end-of-the-year STAR Maths tests were statistically significant

( $p < 0.001$ ). The average number of months between testing occasions was about 6.2. As STAR Maths was vertically scaled, an overall predictive validity coefficient was computed, and found to be 0.82. Within-year estimates had a median of about 0.67, and a range from 0.54 in year 9 to 0.76 in year 3.

**Table 8: Predictive Validity between Pre-test and Post-test Results and the Average Months between Tests**

Year	Sample Size	Correlation	Avg. Months
2	44	0.61	6.4
3	29	0.76	7.0
4	64	0.64	6.4
5	75	0.74	7.0
6	38	0.69	6.8
7	71	0.62	6.4
8	76	0.75	6.3
9	44	0.54	6.0
Overall	441	0.82	6.5

## Validity Evidence from the US

During the course of STAR Maths norming in the US, mathematics achievement data were received for more than 19,000 pupils (9,000 pupils during the STAR Maths 1.x norming in 1998 and an additional 10,000 pupils during the STAR Maths 2.0 norming in 2001). The standardised tests included a variety of well-established instruments including the California Achievement Test (CAT), the Comprehensive Test of Basic Skills (CTBS), the Iowa Tests of Basic Skills (ITBS), the Metropolitan Achievement Test (MAT), the Stanford Achievement Test (SAT-9), the Northwest Evaluation Association (NWEA) Levels Test, and TerraNova, as well as several state assessments from Connecticut, Delaware, Florida, Georgia, Kentucky, Indiana, Illinois, Maryland, Michigan, Mississippi, New York, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Texas, Virginia and Washington. The extent that the STAR Maths test correlates with these tests provides support for its construct validity. That is, strong and positive correlations between STAR Maths and these other instruments provide support for the claim that STAR Maths effectively measures mathematics achievement.

Tables 9–12 present the correlation coefficients between the scores on the STAR Maths test and each of the other test instruments for which data were received. Tables 9 and 11 present validity coefficients for grades 1–6, and Tables 10 and 12 present the validity coefficients for grades 7–12. The bottom of each table presents a grade-by-grade summary, including the total number of pupils for whom test data were available, the number of validity coefficients for that grade and the average value of the validity coefficients. The averages were quite consistent from grade to grade, ranging from 0.47 to 0.79, with a median validity of 0.63.

While these validity coefficients are high, they probably underestimate the actual correlations between the STAR Maths test and the other standardised tests of Mathematics achievement. The actual relationship between the STAR Maths test and these other tests is probably higher than these estimates indicate, as most of these estimates are based on data from intact classrooms, and some restriction of the range of mathematics achievement is to be expected with scores from intact classrooms. Range restriction is well known to attenuate correlation coefficients, as are transcription errors and other clerical errors. Although several safeguards to reduce sources of error were implemented, this procedure was not immune to data collection errors.

The process of establishing the validity of a test is laborious, and it usually takes a significant amount of time. As a result, the validation of the STAR Maths test is an ongoing activity, with the goal of establishing evidence of the test's validity for a variety of settings and pupils. STAR Maths users who collect relevant data are encouraged to contact Renaissance Learning.

Since correlation coefficients are available for many different test editions, forms and administration dates, many of the tests have several validity coefficients associated with them. Where test data quality could not be verified, and when sample size was very small, those data were omitted from the tabulations. Correlations were computed separately on tests according to the unique combination of test edition/form and time when testing occurred. Testing data for other standardised tests administered prior to spring 1998 were excluded from the validity analyses.

In general, these correlation coefficients reflect very well on the validity of the STAR Maths test as a tool for placement in Mathematics. In fact, the correlations are similar in magnitude to the validity coefficients of these measures with each other. These validity results, combined with the supporting evidence of reliability and minimization of SEM estimates for the STAR Maths test, provide quantitative demonstration of how well this innovative Mathematics achievement assessment performs.

**Table 9: Concurrent Validity—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered in Spring 2002, Grades 1–6<sup>a</sup>**

Test	Version	Date	Score	1		2		3		4		5		6	
				n	r	n	r	n	r	n	r	n	r	n	r
California Achievement Test (CAT)															
CAT	5th Ed.	S 02	NCE	–	–	–	–	17	0.50*	–	–	–	–	–	–
Comprehensive Test of Basic Skills (CTBS)															
CTBS	A-13	S 02	SS	–	–	–	–	–	–	–	–	21	0.66*	–	–
CTBS		S 02	NCE	–	–	–	–	–	–	–	–	–	–	32	0.65*
Iowa Tests of Basic Skills (ITBS)															
ITBS	Form A	S 02	NCE	–	–	–	–	–	–	50	0.66*	79	0.72*	–	–
ITBS	Form K	S 02	SS	–	–	–	–	–	–	–	–	–	–	70	0.69*
ITBS	Form L	S 02	NCE	–	–	7	0.78*	23	0.57*	17	0.70*	21	0.66*	–	–
ITBS	Form M	S 02	NCE	14	0.56*	11	0.58	–	–	–	–	–	–	–	–
ITBS	Form M	S 02	SS	–	–	–	–	17	0.72*	–	–	–	–	–	–
McGraw Hill Mississippi/Criterion Referenced (McGraw)															
McGraw		S 02	SS	–	–	–	–	–	–	–	–	44	0.73*	–	–
Metropolitan Achievement Test (MAT)															
MAT	6th Ed.	S 02	NCE	69	0.55*	–	–	–	–	–	–	–	–	–	–
MAT	8th Ed.	S 02	SS	–	–	–	–	–	–	38	0.83*	–	–	–	–
Mississippi Curriculum Test (CTB)															
CTB	Miss	S 02	SS	–	–	–	–	–	–	10	0.62	–	–	–	–
North Carolina End of Grade (NCEOG)															
NCEOG		S 02	NCE	–	–	–	–	70	0.60*	–	–	–	–	–	–
NCEOG		S 02	SS	–	–	–	–	62	0.73*	–	–	–	–	–	–
Oregon State Assessment (Oregon)															
Oregon		S 02	SS	–	–	–	–	–	–	73	0.65*	–	–	–	–
Pennsylvania System of School Assessment (PSSA)															
PSSA		S 02	SS	–	–	–	–	–	–	–	–	–	–	62	0.76*

**Table 9: Concurrent Validity—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered in Spring 2002, Grades 1–6<sup>a</sup> (Continued)**

Test	Version	Date	Score	1		2		3		4		5		6	
				n	r	n	r	n	r	n	r	n	r	n	r
Stanford Achievement Test (Stanford)															
SAT-9		S 02	NCE	–	–	113	0.56*	39	0.83*	46	0.54*	103	0.70*	49	0.65*
SAT-9		S 02	SS	20	0.76*	16	0.68*	18	0.59*	19	0.57*	71	0.49*	84	0.62*
TerraNova															
TerraNova		S 02	NCE	7	0.66	14	0.46	125	0.68*	18	0.67*	17	0.79*	15	0.64*
Summary															
Grade(s)	All			1	2	3	4	5	6						
Number of pupils	1,581			110	161	371	271	356	312						
Number of coefficients	38			4	5	8	8	7	6						
Average validity				0.63	0.61	0.65	0.66	0.68	0.67						
Overall average	0.65														

a. (n) = Sample size. Asterisks (\*) denote correlation coefficients that are statistically significant at the 0.05 level.

**Table 10: Concurrent Validity—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered in Spring 2002, Grades 7–12<sup>a</sup>**

Test	Version	Date	Score	7		8		9		10		11		12	
				n	r	n	r	n	r	n	r	n	r	n	r
Florida Comprehensive Assessment Test (FCAT)															
FCAT		S 02	NCE	–	–	–	–	–	–	51	0.64*	57	0.66*	38	0.75*
Iowa Tests of Basic Skills (ITBS)															
ITBS	Form M	S 02	SS	37	0.40*	–	–	–	–	–	–	–	–	–	–
Michigan Comprehensive Assessment Test (MCAS)															
MCAS		S 02	SS	–	–	–	–	–	–	–	–	112	0.66*	–	–

**Table 10: Concurrent Validity—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered in Spring 2002, Grades 7–12<sup>a</sup> (Continued)**

Test	Version	Date	Score	7		8		9		10		11		12	
				n	r	n	r	n	r	n	r	n	r	n	r
New Standards Reference Mathematics Exam (Rhode Island) (NSRME)															
NSRME	RI	S 02	SS	–	–	–	–	–	–	–	–	67	0.67*	9	0.66
Ohio Proficiency Test (Ohio)															
Ohio		S 02	SS	–	–	–	–	23	0.67*	26	0.40*	24	0.77*	24	0.69*
Otis Lennon School Ability Test (OLSAT)															
OLSAT		S 02	NCE	–	–	–	–	–	–	12	0.36	13	0.91*	6	0.72
Palmetto Achievement Challenge Test 2001 (PACT)															
PACT	2001	S 02	SS	–	–	161	0.72*	–	–	–	–	–	–	–	–
Stanford Achievement Test (Stanford)															
SAT-9		S 02	NCE	–	–	–	–	–	–	–	–	–	–	15	0.54*
SAT-9		S 02	SS	59	0.57*	9	0.85*	–	–	–	–	–	–	–	–
Texas Assessment of Academic Skills, 2001 (TAAS)															
TAAS	2001	S 02	TLI	–	–	–	–	163	0.69*	–	–	–	–	–	–
Summary															
Grade(s)	All			7		8		9		10		11		12	
Number of pupils	906			96		170		186		89		273		92	
Number of coefficients	19			2		2		2		3		5		5	
Average validity				0.49		0.79		0.68		0.47		0.73		0.67	
Overall average	0.65														

a. (n) = Sample size. Asterisks (\*) denote correlation coefficients that are statistically significant at the 0.05 level.

**Table 11: Other External Validity Data—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered Prior to Spring 2002, Grades 1–6<sup>a</sup>**

Test	Version	Date	Score	1		2		3		4		5		6	
				n	r	n	r	n	r	n	r	n	r	n	r
Achievement Level Test (RIT)															
RIT		F 01	SS	–	–	–	–	–	–	–	–	–	–	150	0.69*
California Achievement Test (CAT)															
CAT	5th Ed.	S 01	SS	–	–	–	–	46	0.52*	–	–	–	–	–	–
Cognitive Abilities Test (CogAT)															
CogAT		F 00	SS	–	–	–	–	41	0.61*	–	–	–	–	–	–
CogAT		F 01	SS	–	–	45	0.73*	–	–	–	–	–	–	–	–
Comprehensive Test of Basic Skills (CTBS)															
CTBS	4th Ed.	S 01	GE	–	–	–	–	–	–	43	0.67*	–	–	–	–
CTBS	A-13	S 00	NCE	–	–	–	–	–	–	65	0.60*	–	–	–	–
CTBS	A-13	S 00	SS	–	–	–	–	–	–	–	–	44	0.70*	–	–
CTBS	A-13	S 01	GE	–	–	–	–	–	–	–	–	–	–	56	0.69*
CTBS	A-13	S 01	NCE	–	–	–	–	–	–	–	–	67	0.72*	–	–
CTBS	A-13	S 01	SS	–	–	–	–	–	–	42	0.61*	–	–	–	–
Connecticut Mastery Test (Conn)															
Conn	2nd	F 00	SS	–	–	–	–	–	–	–	–	35	0.51*	–	–
Conn	3rd	F 01	SS	–	–	–	–	–	–	42	0.64*	–	–	27	0.52*
Des Moines Public School (Grade 2 pretest) (DMPS)															
DMPS		F 01	NCE	–	–	25	0.76*	–	–	–	–	–	–	–	–
Educational Development Series (EDS)															
EDS	13C	S 01	GE	–	–	–	–	30	0.69*	–	–	–	–	–	–
EDS	14C	S 00	GE	–	–	–	–	–	–	32	0.44*	–	–	–	–
EDS	15C	F 01	GE	–	–	–	–	–	–	–	–	37	0.68*	–	–
Florida Comprehensive Assessment Test (FCAT)															
FCAT		S 01	NCE	–	–	–	–	–	–	–	–	73	0.65*	–	–

**Table 11: Other External Validity Data—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered Prior to Spring 2002, Grades 1–6<sup>a</sup> (Continued)**

Test	Version	Date	Score	1		2		3		4		5		6	
				n	r	n	r	n	r	n	r	n	r	n	r
Iowa Tests of Basic Skills (ITBS)															
ITBS	Form A	S 01	NCE	–	–	–	–	73	0.45*	78	0.65*	–	–	–	–
ITBS	Form A	F 01	NCE	–	–	–	–	25	0.41*	25	0.35	23	0.33	86	0.81*
ITBS	Form A	F 01	SS	–	–	–	–	–	–	–	–	–	–	73	0.64*
ITBS	Form K	F 00	SS	–	–	–	–	–	–	–	–	–	–	20	0.92*
ITBS	Form K	S 01	NCE	–	–	101	0.67*	74	0.64*	31	0.25	11	0.58	31	0.62*
ITBS	Form K	F 01	NCE	–	–	–	–	10	0.78*	16	0.78*	9	0.54	18	0.63*
ITBS	Form K	F 01	SS	–	–	–	–	–	–	–	–	75	0.77*	68	0.71*
ITBS	Form L	S 01	NCE	–	–	–	–	13	0.50	46	0.81*	13	0.73*	–	–
ITBS	Form L	S 01	SS	–	–	–	–	–	–	11	0.81*	–	–	–	–
ITBS	Form L	F 01	NCE	–	–	–	–	–	–	–	–	69	0.66*	–	–
ITBS	Form M	S 99	NCE	–	–	–	–	–	–	–	–	–	–	19	0.68*
ITBS	Form M	S 00	NCE	–	–	–	–	–	–	–	–	28	0.65*	–	–
ITBS	Form M	S 01	NCE	–	–	19	0.81*	–	–	43	0.78*	–	–	–	–
ITBS	Form M	S 01	SS	–	–	–	–	47	0.39*	32	0.55*	–	–	–	–
ITBS	Form M	F 01	NCE	5	0.88*	–	–	–	–	15	0.82*	–	–	–	–
McGraw Hill Mississippi/Criterion Referenced (McGraw)															
McGraw		S 01	SS	–	–	–	–	–	–	–	–	121	0.52*	–	–
Metropolitan Achievement Test (MAT)															
MAT	7th Ed.	F 01	NCE	–	–	–	–	–	–	–	–	–	–	15	0.84*
Michigan Education Assessment Program (MEAP)															
MEAP		S 01	SS	–	–	–	–	–	–	–	–	88	0.72*	–	–
Multiple Assessment Series (Primary Grades) (Multiple)															
Multiple		S 01	NCE	–	–	14	0.52	19	0.54*	–	–	–	–	–	–
New York State Math Assessment (NYSMA)															
NYSMA		S 01	SS	–	–	–	–	–	–	–	–	50	0.79*	–	–

**Table 11: Other External Validity Data—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered Prior to Spring 2002, Grades 1–6<sup>a</sup> (Continued)**

Test	Version	Date	Score	1		2		3		4		5		6	
				n	r	n	r	n	r	n	r	n	r	n	r
North Carolina End of Grade (NCEOG)															
NCEOG		F 01	SS	–	–	–	–	85	0.57*	–	–	–	–	–	–
Northwest Evaluation Association Levels Test (NWEA)															
NWEA		S 01	NCE	–	–	–	–	–	–	–	–	83	0.81*	64	0.78*
NWEA		F 01	NCE	–	–	–	–	50	0.56*	49	0.54*	99	0.70*	–	–
Ohio Proficiency Test (Ohio)															
Ohio		S 01	SS	–	–	–	–	113	0.65*	–	–	–	–	–	–
Stanford Achievement Test (Stanford)															
SAT-9		S 99	SS	–	–	–	–	–	–	–	–	55	0.65*	–	–
SAT-9		S 00	SS	–	–	–	–	–	–	–	–	–	–	15	0.50
SAT-9		F 00	NCE	–	–	–	–	17	0.84*	20	0.83*	–	–	–	–
SAT-9		F 00	SS	–	–	–	–	–	–	–	–	–	–	46	0.58*
SAT-9		S 01	NCE	–	–	–	–	43	0.69*	–	–	50	0.38*	–	–
SAT-9		S 01	SS	64	0.52*	–	–	–	–	58	0.41*	52	0.58*	51	0.65*
SAT-9		F 01	SS	–	–	–	–	–	–	90	0.54*	32	0.67*	24	0.57*
Tennessee Comprehensive Assessment Program, 2001 (TCAP)															
TCAP	2001	S 01	SS	–	–	–	–	–	–	–	–	48	0.56*	–	–
TerraNova															
TerraNova		S 00	NCE	–	–	–	–	–	–	–	–	–	–	43	0.60*
TerraNova		S 00	SS	–	–	–	–	–	–	–	–	11	0.61*	–	–
TerraNova		F 00	SS	–	–	–	–	–	–	–	–	108	0.62*	–	–
TerraNova		S 01	NCE	–	–	–	–	–	–	–	–	69	0.40*	85	0.62*
TerraNova		S 01	SS	–	–	–	–	–	–	104	0.50*	62	0.59*	131	0.71*
TerraNova		F 01	NCE	–	–	58	0.38*	63	0.56*	70	0.74*	85	0.61*	–	–
Test of New York State Standards (TONYSS)															
TONYSS		S 01	SS	–	–	–	–	55	0.75*	68	0.47*	–	–	–	–

**Table 11: Other External Validity Data—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered Prior to Spring 2002, Grades 1–6<sup>a</sup> (Continued)**

Test	Version	Date	Score	1		2		3		4		5		6	
				n	r	n	r	n	r	n	r	n	r	n	r
Texas Assessment of Academic Skills (TAAS)															
TAAS	2001	S 01	SS	–	–	–	–	–	–	78	0.52*	–	–	–	–
TAAS	2001	S 01	TLI	–	–	–	–	–	–	–	–	–	–	82	0.42*
Virginia Standards of Learning (Virginia)															
Virginia		S 00	SS	–	–	–	–	–	–	–	–	24	0.73*	–	–
Washington Assessment of Student Learning (Wash)															
Wash		S 00	SS	–	–	–	–	–	–	–	–	–	–	90	0.54*
Wide Range Achievement Test (WRAT)															
WRAT III		F 01	NCE	–	–	–	–	–	–	44	0.32*	44	0.66*	–	–
Summary															
Grade(s)	All			1	2	3	4	5	6						
Number of pupils	4,996			69	262	804	1,102	1,565	1,194						
Number of coefficients	98			2	6	17	23	29	21						
Average validity				0.70	0.65	0.60	0.59	0.62	0.65						
Overall average	0.62														

a. (n) = Sample size. Asterisks (\*) denote correlation coefficients that are statistically significant at the 0.05 level.

**Table 12: Other External Validity Data—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered Prior to Spring 2002, Grades 7–12<sup>a</sup>**

Test	Version	Date	Score	7		8		9		10		11		12	
				n	r	n	r	n	r	n	r	n	r	n	r
American College Testing Program (ACT)															
ACT		F 01	NCE	–	–	–	–	–	–	–	–	–	–	26	0.87*
California Achievement Tests (CAT)															
CAT	5th Ed.	F 01	NCE	–	–	–	–	64	0.73*	–	–	–	–	–	–
CAT	5th Ed.	F 01	SS	170	0.54*	–	–	–	–	–	–	–	–	–	–
Comprehensive Test of Basic Skills (CTBS)															
CTBS	4th Ed.	S 00	SS	67	0.67*	75	0.73*	–	–	–	–	–	–	–	–
CTBS	A-13	S 00	SS	–	–	31	0.65*	–	–	–	–	–	–	–	–
CTBS	A-13	S 01	SS	23	0.82*	–	–	–	–	48	0.63*	–	–	–	–
Delaware Student Testing Program (DSTP)															
DSTP		S 01	SS	–	–	–	–	94	0.27*	–	–	–	–	–	–
Differential Aptitude Tests (DAT)															
DAT	Level 1	F 01	NCE	–	–	–	–	41	0.70*	–	–	–	–	–	–
Explore Tests (Explore)															
Explore		F 01	NCE	–	–	64	0.54*	–	–	–	–	–	–	–	–
Georgia High School Graduation Test (Georgia)															
Georgia		S 01	NCE	–	–	–	–	–	–	–	–	–	–	23	0.71*
Indiana Statewide Testing for Educational Progress (ISTEP)															
ISTEP		F01	NCE	–	–	–	–	51	0.57*	22	0.58*	–	–	–	–
Iowa Tests of Basic Skills (ITBS)															
ITBS	Form A	F 01	SS	66	0.71*	–	–	–	–	–	–	–	–	–	–
ITBS	Form K	S 01	NCE	73	0.80*	18	0.52*	–	–	–	–	–	–	–	–
ITBS	Form K	F 01	NCE	6	0.72	14	0.69*	–	–	–	–	–	–	–	–
ITBS	Form L	S 01	NCE	36	0.74*	32	0.53*	–	–	19	0.67*	32	0.84*	–	–
ITBS	Form M	S 99	NCE	–	–	5	0.89*	–	–	–	–	11	0.80*	–	–
ITBS	Form M	S 00	NCE	–	–	–	–	–	–	9	0.94*	–	–	–	–
ITBS	Form M	S 01	NCE	49	0.52*	48	0.51*	–	–	–	–	–	–	–	–

**Table 12: Other External Validity Data—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered Prior to Spring 2002, Grades 7–12<sup>a</sup> (Continued)**

Test	Version	Date	Score	7		8		9		10		11		12	
				n	r	n	r	n	r	n	r	n	r	n	r
Kentucky Core Content Test (KCCT)															
KCCT		S 01	NCE	–	–	–	–	45	0.43*	–	–	–	–	–	–
Maryland High School Placement Test (Maryland)															
Maryland		S 01	NCE	–	–	–	–	47	0.60*	–	–	–	–	–	–
McGraw Hill Mississippi/Criterion Referenced (McGraw)															
McGraw		S 01	SS	–	–	–	–	73	0.56*	–	–	–	–	–	–
Metropolitan Achievement Test (MAT)															
MAT	7th Ed.	F 01	NCE	5	0.80	11	0.82*	–	–	–	–	–	–	–	–
North Carolina End of Grade Tests (NCEOG)															
NCEOG		S 01	SS	–	–	177	0.59*	–	–	–	–	–	–	–	–
Oklahoma School Testing Program Core Curriculum Tests (Oklahoma)															
Oklahoma		S 01	SS	–	–	–	–	26	0.67*	–	–	–	–	–	–
Oregon State Assessment (Oregon)															
Oregon		S 01	NCE	46	0.49*	45	0.53*	–	–	–	–	–	–	–	–
PLAN															
PLAN		F 99	SS	–	–	–	–	–	–	–	–	–	–	10	0.42
PLAN		F 00	SS	–	–	–	–	–	–	–	–	40	0.28	–	–
PLAN		F 01	NCE	–	–	–	–	–	–	63	0.61*	–	–	–	–
Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT)															
PSAT/NMSQT	NMSQT	F 00	NCE	–	–	–	–	–	–	–	–	–	–	37	0.63*
PSAT/NMSQT	NMSQT	F 01	NCE	–	–	–	–	–	–	–	–	72	0.64*	–	–
Stanford Achievement Test (Stanford)															
SAT-9		S 98	NCE	11	0.84*	–	–	–	–	–	–	–	–	–	–
SAT-9		S 99	NCE	14	0.71*	–	–	–	–	–	–	–	–	–	–
SAT-9		F 00	SS	–	–	45	0.85*	–	–	–	–	–	–	–	–
SAT-9		S 01	NCE	45	0.71*	105	0.81*	11	0.69*	–	–	–	–	–	–
SAT-9		S 01	SS	54	0.76*	109	0.69*	19	0.27	77	0.59*	67	0.76*	71	0.65*
SAT-9		F 01	SS	104	0.84*	–	–	–	–	–	–	–	–	–	–

**Table 12: Other External Validity Data—STAR Maths 2.0 Correlation Coefficients (r) with External Tests Administered Prior to Spring 2002, Grades 7–12<sup>a</sup> (Continued)**

Test	Version	Date	Score	7		8		9		10		11		12	
				n	r	n	r	n	r	n	r	n	r	n	r
TerraNova															
TerraNova		S 99	NCE	35	0.61*	47	0.62*	–	–	–	–	–	–	–	–
TerraNova		S 00	SS	18	0.73*	–	–	–	–	–	–	–	–	–	–
TerraNova		S 01	NCE	17	0.29	17	0.52*	–	–	–	–	–	–	–	–
TerraNova		S 01	SS	–	–	99	0.74*	–	–	–	–	–	–	–	–
TerraNova		F 01	SS	–	–	38	0.74*	–	–	–	–	–	–	–	–
Test of Achievement Proficiency (TAP)															
TAP		F 01	NCE	–	–	–	–	8	0.70	7	0.70	–	–	–	–
Texas Assessment of Academic Skills, 2001 (TAAS)															
TAAS	2001	S 01	SS	66	0.44*	69	0.33*	–	–	–	–	–	–	–	–
Virginia Standards of Learning (Virginia)															
Virginia		S 00	SS	25	0.71*	–	–	–	–	–	–	–	–	–	–
Summary															
Grade(s)	All			7		8		9		10		11		12	
Number of pupils	3,066			930		1,049		479		245		222		141	
Number of coefficients	66			20		19		11		7		5		4	
Average validity				0.67		0.65		0.56		0.67		0.66		0.60	
Overall average	0.64														

a. (n) = Sample size. Asterisks (\*) denote correlation coefficients that are statistically significant at the 0.05 level.

## Types of Test Scores

After pupils have tested with STAR Maths, the software uses their test results to determine three types of test scores that express pupils' mathematics performance: Scale Score, Estimated National Curriculum Level and Criterion-Referenced Score.

### Scale Score (SS)

STAR Maths 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, all STAR Maths test scores are converted to a common scale, creating Scale Scores. The STAR Maths 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, using a linear transformation to make all scores positive integers, the Rasch ability scores are converted to STAR Maths Scale Scores. STAR Maths Scale Scores range from 1 to 1400.

STAR Maths Scale Scores are expressed on the same scale across all versions of the software. STAR Maths Scale Scores provide a single scale for measuring the mathematics achievement of pupils from years 2–13. In addition, STAR Maths norm-referenced scores are derived from the within-grade distributions of Scale Scores in the STAR Maths 2.0 US norms group.

### Estimated National Curriculum Level–Maths (Est. NCL)

The Estimated National Curriculum Level (Est. NCL) in Mathematics is an estimate of a pupil's standing on the National Curriculum based on his or her STAR Maths performance. This score is an approximation based on the demonstrated relationship between STAR Maths scale scores and teachers' judgments through their teacher assessment (TA) of pupil's obtained skills. It should not be taken to be the pupil's actual national curriculum level, but rather an estimate of the level at which the child is most likely performing. Stating this another way, the Est. NCL from STAR Maths is an estimate of the individual's standing in the national curriculum framework based on a modest number of STAR Maths 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, an Est. NCL of 4c would indicate that an individual is estimated to have just obtained level 4, while another pupil with 4a is estimated to be approaching level 5.

It is sometimes difficult to determine whether a pupil is in the top of one level (for instance, 4a) or just beginning the next higher 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 indicated by concatenation of

the contiguous levels and sublevel categories. For instance, a pupil whose skills appear to range between levels 4 and 5, indicating they are probably starting to transition from one level to the next, would obtain an 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.

## Criterion-Referenced Scores

Criterion-referenced scores describe a pupil'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 percentage-correct score, which expresses what proportion of test questions the pupil can answer correctly in the content domain.

An example of a criterion-referenced classification is a proficiency category on a standards-based assessment: the pupil 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 Numeration and Computation mastery classification charts in the Diagnostic Report are criterion-referenced.

# APPENDIX

---

## References

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA.

## 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

Copyright © 2007, Renaissance Learning, Inc. All Rights Reserved. Printed in the United States of America.

Renaissance Learning, the Renaissance Learning logo, the STAR logo, and STAR Maths are trademarks of Renaissance Learning, Inc. and its subsidiaries, registered, common law, or pending registration, in the United States and in other countries.



32 Harbour Exchange Square  
London E14 9GE

Tel: 020 7184 4000  
Fax: 020 7538 2625  
Email: [info@renlearn.co.uk](mailto:info@renlearn.co.uk)  
Website: [www.renlearn.co.uk](http://www.renlearn.co.uk)