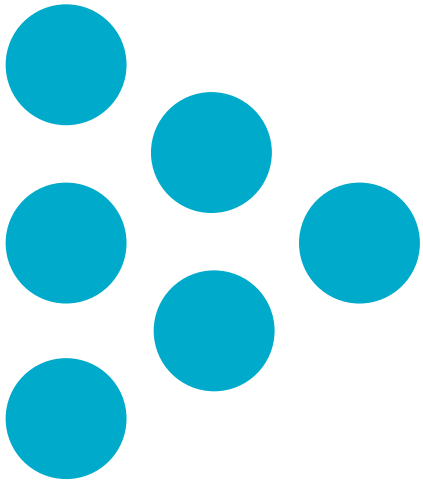
NFER Classroom

Technical Report

Technical information for NFER Tests in reading and mathematics

Suite 2 (Summer)

National Foundation for Educational Research (NFER)



Technical Information for NFER Tests in reading and mathematics
Suite 2 (Summer)

Centre for Assesment

Published in September 2020

By the National Foundation for Educational Research,
The Mere, Upton Park, Slough, Berkshire SL1 2DQ

www.nfer.ac.uk

ISBN: 978-1-912596-04-1

How to cite this publication:

Centre for Assessment (2020). *Technical Information for NFER Tests in reading and mathematics Suite 2 (Summer).* Slough: NFER.

Contents

1 Introduction

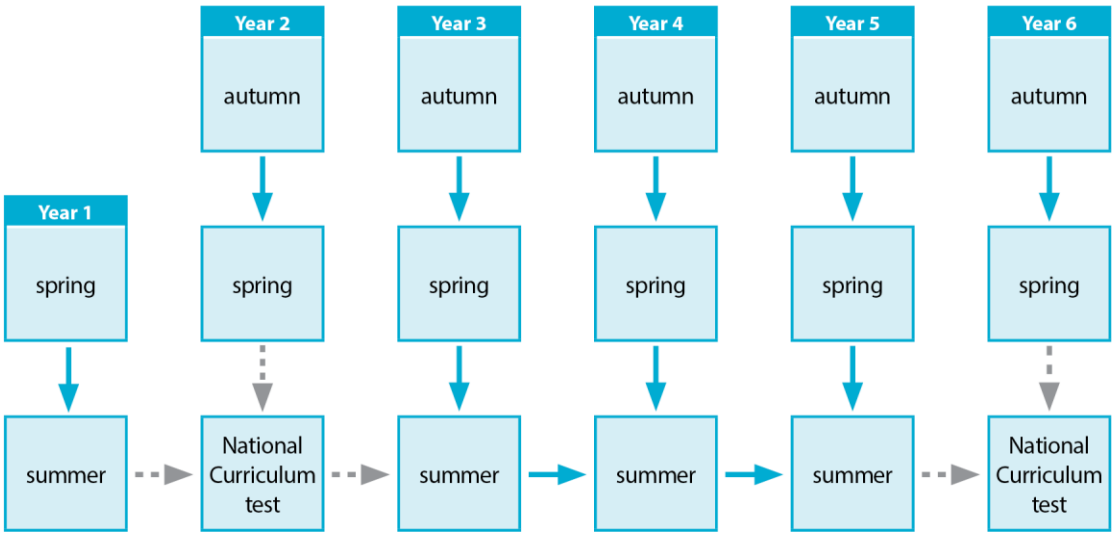
This manual has been published for transparency and to demonstrate the quality of the NFER Tests, so that readers can understand the rigorous development process and appraise the trial data which supports the published materials. It is intended to be of interest to an audience with knowledge of assessment, for example those who develop tests or those who take an assessment lead in schools.

2 The NFER Tests

Following the introduction of the new national curriculum in 2014 and the abolition of the eight-level scale of assessment, NFER developed a new suite of tests to help inform teacher assessment. The design of these tests reflects the changes to the model of statutory assessment used since 2016 and they have been standardised with a large nationally representative sample of pupils who have been taught the new curriculum for at least a year and at a time point in the school year which matches the intended use of the tests.

The suite has been expanded across the years and now consists of a series of termly tests for use in Year 1 through to Year 6. The diagram below shows the current extent of the suite and indicates the possible pathways through which pupil progress can be monitored. In addition to the tests themselves, NFER Tests users also have access to the NFER Tests Analysis Tool. This tool can be used to record pupils' marks and results across the whole suite of tests and therefore enables progress to be monitored between terms and across years. The arrows on the figure below show some of the stages between which progress can be monitored. However, the tool allows comparison between any two terms and also provides an indication of pupil performance across the different skills that make up the assessment. Looking at the way in which a pupil is progressing relative to their peer group within the school and on a national scale, will help teachers to identify pupils who may be in need of additional support in order to make more progress. (The term 'item' is used within test development to refer to a numbered question within the test.)

Although the suite of tests is extensive, this report pertains only to the reading and mathematics summer tests for Years 3 to 5 as these tests were developed in tandem.



3 Early development of texts and items

Following the initial development of texts / contexts and items by the researchers at NFER, qualitative trialling was conducted at a variety of primary schools. Qualitative trialling involves discussing the texts and/or items with small groups of pupils and gathering information on how these can be improved. This provides early feedback on the appropriateness of the texts and items, contributes to an informed review of the materials and influences the selection of items in preparation for the standardisation trial.

Teacher feedback is very important in the development of NFER Tests. Not only is teacher input gathered on the early versions of the materials during informal trialling but it is also collected through a questionnaire completed by teachers taking part in the large scale standardisation trial. This questionnaire gathers teacher feedback on different aspects of the tests; this information is very useful in refining the materials and informing the selection of items that comprise the final tests.

In addition to feedback from teachers, the materials were reviewed by inclusion and subject experts. This allows us to ensure that, as far as possible, the tests are appropriate for the pupils who will be taking them.

4 Standardisation sample characteristics

The NFER Tests in reading and mathematics suite 2 were standardised in June 2015 with a sample of schools from across England. Around 4500 pupils participated in the trial of the materials for each subject (reading and mathematics).

The standardisation trial has several purposes. Firstly, it provides item level data from which we can discern exactly how each pupils have performed on each question. This enables us to eliminate items which pupils have misunderstood or not completed as expected. This may be because of imprecise or misleading wording or some other source of misunderstanding. Additionally it allows us to remove from the item pool any items that are too hard or too easy, and to select a final set of items which present an appropriate range of difficulty overall.

A second purpose of the standardisation trial is to refine mark schemes. This is done by selecting exemplar responses that pupils give to items during the trial to refine and clarify the marking points. In addition, for responses on the borderline, knowing the proportions of pupils that have given certain types of responses and the associated ability of these pupils, we can also make final decisions as to which responses may be credited or not.

Of course, the standardisation trial is also used to collect data which enables us to calculate the standardised scores provided in the teacher guide and available in the Analysis Tool. These standardised scores enable schools to compare the performance of each child against the performance of other children nationally or within their own school. When standardising a test it is important to ensure that the sample of schools taking the test is representative of the national school population. In order to select the sample, all schools in England were divided into separate groups, called strata, based on their characteristics. This was carried out for several characteristics (stratifiers) including school type. In this stratifier, the strata are: primary/combined schools, junior schools, middle schools and independent schools. A random sample is then selected to match the proportions of schools nationally in each stratum, a process known as 'stratified sampling'. The standardisation sample for these tests was stratified according to the following characteristics:

- KS2 overall performance band 2015 (average point score)
- Region: government office region.

When a standardisation sample is selected it is necessary to ensure that the percentage of schools in each of the groups (strata) reflects the national picture. For example, if nationally 84 per cent of schools are categorised as primary schools then this should be mirrored in the sample (i.e. around 84 per cent of the sample should be primary schools). In order to ensure the characteristics of the schools included in the standardisation sample were representative nationally, school level characteristics were compared with the national population and chi-squared significance tests[1] were conducted. The achieved sample representations across the above characteristics are shown and compared with the national population in Tables 1 and 2. The gender breakdown of the

[1] A chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories.

sample is shown in Table 3. All data relates to the standardisation of 2015. The samples were representative of the national population at the school level.

Table 1: Representation of the sample at school level – Years 3, 4 and 5 reading

		Population		Sample	
		Number	**%**	**Number**	**%**
KS2 overall performance band 2013 (average point score)	Lowest 20%	2616	20	7	10
	2nd lowest 20%	2418	18	12	18
	Middle 20%	2639	20	14	21
	2nd highest 20%	2447	19	18	26
	Highest 20%	2957	23	17	25
Government Office Region	North East	747	5	2	3
	North West/Merseyside	2467	16	8	11
	Yorkshire & The Humber	1715	11	10	14
	East Midlands	1512	10	7	9
	West Midlands	1621	10	12	16
	Eastern	1729	11	9	12
	London	1919	12	6	8
	South East	2400	15	11	15
	South West	1804	11	9	12
Total		**15914**	**101**	**74**	**100**

Since percentages are rounded to the nearest integer, they may not always sum to 100.

The Year 3, 4 and 5 reading sample is representative of the national population at the school level. Any differences between the population and the achieved sample are small and are not statistically significant.

Table 2: Representation of the sample at school level – Year 3, 4 and 5 mathematics (A, T1 and T2)

		Population		Sample	
		Number	**%**	**Number**	**%**
KS2 overall performance band 2013 (average point score)	Lowest 20%	2616	20	7	10
	2nd lowest 20%	2418	19	12	18
	Middle 20%	2639	20	13	19
	2nd highest 20%	2447	19	18	27
	Highest 20%	2957	23	17	25
Total		**13077**	**101**	**67**	**99**
Government Office Region	North East	747	5	2	3
	North West/ Merseyside	2467	16	8	11
	Yorkshire & The Humber	1715	11	9	13
	East Midlands	1512	10	6	8
	West Midlands	1621	10	12	17
	Eastern	1729	11	9	13
	London	1919	12	6	8
	South East	2400	15	11	15
	South West	1804	11	9	13

		Population		Sample	
		Number	%	Number	%
Total		**15914**	**101**	**72**	**101**

Since percentages are rounded to the nearest integer, they may not always sum to 100.

The Year 3, 4 and 5 mathematics sample is representative of the national population at the school level. Any differences between the population and the achieved sample are small and are not statistically significant.

Table 3: Representation of the sample at school level: gender

	Population	
	Number	**%**
Female	2281	49
Male	2370	51

In terms of gender, all the year group samples were representative of the national population at pupil level.

5 Whole test functioning

The following tables provide information on the overall performance (or "functioning") of each test separately by year group. An explanation of each measure listed in the tables is provided below.

Standardisation sample *(n)*: A standardised test is one that has been trialled with a nationally representative sample of pupils. The size of the pupil sample is important if you are benchmarking pupils against attainment nationally as larger samples give more accurate scores. The tables below show that the tests have been standardised on sufficiently large samples and therefore provide accurate standardised scores.

Reliability *(Cronbach's alpha)*: Cronbach's alpha is a statistical measure of internal consistency, which is an aspect of reliability. It indicates the strength of the relationship between all the items in the test. It is a form of 'split-half' reliability, which means if you split the test into two similar sub-tests it tells you how consistent the scores on the two halves of the test would be. The values produced are a form of correlation and can range from 0 to 1; values above 0.8 are usually considered to indicate good reliability. Cronbach's alpha was calculated for each test and the figures in the table below show that all tests were found to have good levels of reliability.

Maximum score: The maximum score is the available number of marks on each test.

Mean and Median: The mean and the median are both measures of central tendency; they give an indication of the average value of a distribution of scores.

The mean is the arithmetic average of a group of scores; that is, the scores are added up and divided by the total number of scores.

The median is the middle score in a list of scores written in numerical order; it is the score point at which half the scores are greater and half the scores smaller.

Standard deviation (SD): Standard deviation is a measure of the amount of variation or dispersion of a set of data values. Put simply, it is telling you how diverse the scores on this test were. A low SD indicates that the scores tend to be close to the mean, whereas a high SD indicates that the scores are spread out over a large range of values. The SD will, to some extent, be limited by the total number of marks available in each test.

Table 4: Whole test functioning by test: Year 3

	Year 3 reading	Year 3 mathematics arithmetic paper	Year 3 mathematics Test 1	Year 3 mathematics Test 2
Standardisation sample *n*	1530	1514	1514	1513
Reliability (Cronbach's alpha)	0.916	0.926	0.883	0.879
Maximum score	37	30	25	25
Mean	20.13	14.64	12.02	11.66
Median	21.00	15.00	12.00	11.00
Standard deviation	9.00	8.02	6.24	6.03

Table 5: Whole test functioning by test: Year 4

	Year 4 reading	Year 4 mathematics arithmetic paper	Year 4 mathematics Test 1	Year 4 mathematics Test 2
Standardisation sample *n*	1539	1517	1520	1515
Reliability (Cronbach's alpha)	0.917	0.932	0.890	0.905
Maximum score	40	35	30	30
Mean	19.52	17.6	14.46	14.75
Median	19.00	17.00	14.00	14.00
Standard deviation	9.26	9.22	6.98	7.59

Table 6: Whole test functioning by test: Year 5

	Year 5 reading	Year 5 mathematics arithmetic paper	Year 5 mathematics Test 1	Year 5 mathematics Test 2
Standardisation sample *n*	1584	1529	1535	1530
Reliability (Cronbach's alpha)	0.903	0.942	0.917	0.921
Maximum score	46	40	35	35
Mean	20.2	17.45	15.60	14.89
Median	20	16.00	15.00	14.00
Standard deviation	9.5	10.46	8.52	9.23

6 Item level functioning

6.1 Item level statistics

Information about item functioning is available in the NFER Tests Analysis Tool. This is available on the NFER portal for purchasers of the Teacher Guides. It provides an indication of the difficulty of each item so that teachers can see whether an item that their pupils found difficult was also generally difficult for the population or alternatively was completed more easily by the population and therefore performance maybe symptomatic of an underlying misconception or gap in teaching.

6.2 Differential item functioning

During the development of the tests we analysed whether different groups of pupils performed differently on the test items. This was carried out using differential item functioning (DIF) analysis and separate analyses were completed for gender and EAL. DIF identifies particular items for which two groups (e.g. girls and boys) perform differently above and beyond the disparity in their achievement on the test as a whole. This analysis is one way of establishing whether or not there could be any bias in the test items, that is, are there items which potentially discriminate inappropriately against one group of learners? The results of this analysis are important as they demonstrate that performance on these NFER Tests is not related to other factors irrelevant to the skill being tested.

However, it is important to recognise that sometimes there are valid reasons why one group might perform differently to another. Therefore, although the presence of DIF *may* indicate that an item may be biased, it does not necessarily mean that the item is unfair. For example, it is recognised that EAL pupils often perform better on mathematics items using specific technical vocabulary as they are more used to learning words and meanings than native speakers, while native speakers often do better at written '*explain your answer*' items. In reading, there is a tendency for girls, on average, to perform slightly better than boys on items requiring an understanding of character.

A number of items within the tests showed differential item functioning, although it should be noted that similar results may not occur if the materials were trialled with a different sample. The results of the DIF analysis are presented in terms of the severity of any difference in performance relative to that expected given the overall difference on the test as a whole. There are three levels of severity: negligible, medium and large. The greater the severity, the larger the magnitude of the differential performance. Experience suggests that items classified as having 'negligible' DIF have minimal impact on the overall difference in performance.

Where DIF analysis identified items with a significant difference in performance between two groups, the items were reviewed to ensure that there were no specific features of that item that would make it globally biased towards one group or the other (e.g. gender). Given that there are always likely to be items within a test that demonstrate DIF, it is important to ensure that across the test the effect of DIF is largely balanced out. The tables below show DIF performance by gender and EAL is generally balanced. The reasoning papers in the mathematics tests tend to have slightly more items favouring non-EAL pupils than EAL pupils.

Table 7: Differential item functioning by gender in the reading tests

Year	Total number of items	Number of items with no statistically significant DIF	Number of items with a DIF greater than negligible
3	32	26	Girls: 0 Boys: 1
4	34	25	Girls: 1 Boys: 1
5	37	21	Girls: 2 Boys: 3

Table 8: Differential item functioning by gender in the mathematics tests

Year	Total number of items	Number of items with no statistically significant DIF	Number of items with a DIF greater than negligible
3	77	48	Girls: 9 Boys: 7
4	88	39	Girls: 10 Boys: 14
5	103	56	Girls: 6 Boys: 11

Table 9: Differential item functioning by EAL in the reading tests

Year	Total number of items	Number of items with no statistically significant DIF	Number of items with a DIF greater than negligible
3	32	31	EAL: 0 Non-EAL: 1
4	34	26	EAL: 2 Non-EAL: 1
5	37	32	EAL: 4 Non-EAL: 0

Table 10: Differential item functioning by EAL in the mathematics tests

Year	Total number of items	Number of items with no statistically significant DIF	Number of items with a DIF greater than negligible
3	77	64	EAL: 4 Non-EAL: 4
4	88	64	EAL: 11 Non-EAL: 9
5	103	90	EAL: 4 Non-EAL: 9

7 Test outcomes

The following outcomes are available from this suite of tests:

- Raw score – the total number of marks attained by each pupil
- Standardised score
- Age standardised score.

More details of each are available in the relevant teacher guide.

It is worth noting that the scaled score of 100 defined by the Department for Education as the national expectation at the end of Key Stage 2 is **not the same as, nor equivalent to**, a standardised score or age standardised score of 100 on these tests. On NFER Tests, a standardised score or age standardised score of 100 represents the average performance, based on a normal distribution, of the sample of pupils on which the tests were standardised. At the end of Key Stage 2, the DfE's scaled score of 100 represents the 'expected standard' and is not the average.

Standardised scores

Standardised scores enable a comparison to be made between the performance of a pupil and that of a large nationally representative sample who took the same test. Such comparisons can be useful for grouping a class by ability and for identifying those pupils in need of targeted interventions. Standardised scores can be averaged to provide an overview of the performance of the class as a whole.

The average standardised score is set at 100, based on the performance of a nationally representative sample. About two-thirds of pupils will have standardised scores between 85 and 115 and scores within this range can be broadly described as 'average'. Almost all pupils fall within the range 70 to 140. The test is not able to distinguish between pupils performing above or below this range as such pupils are not performing at the level of the test. As reliable standardised scores cannot be obtained outside of this range, they are not produced. In some reports scores outside of the range may be denoted 69 and 141 to enable them to be plotted.

It may be helpful to further divide the average category in which case scores from 85 to 94 inclusive may be classified as 'low average' and scores from 106 to 115 inclusive may be classified as 'high average'. Scores from 85 to 105 remain as 'average

Standardisation score	Description	
70 to 84	Below average	
85 to 94	Low average	All pupils within this group are working at an average standard
95 to 105	Average	
106 to 115	High average	
116 to 140	Above average	

Age standardised scores

Age standardised scores take into account a pupil's age in years and months at the time of sitting a test, in order that his or her performance can be compared with the performance of other pupils the same age in a nationally representative sample. The age standardisation that has been undertaken on the NFER Tests means that these tests can be administered at different time points and comparative information still be obtained.

As with standardised scores, the average age standardised score is set at 100, based on the performance of a nationally representative sample. About two-thirds of pupils will have standardised scores between 85 and 115 and scores within this range can be broadly described as 'average'. Almost all pupils fall within the range 70 to 140. As stated above, the test is not able to distinguish between pupils performing above or below this range as such pupils are not performing at the level of the test. As reliable age standardised scores cannot be obtained outside of this range, they are not produced. In some reports scores outside of the range may be denoted 69 and 141 to enable them to be plotted.

NFER Classroom

Evidence for excellence in education

Public

The Mere, Upton Park, Slough, Berks SL1 2DQ
T: +44 (0)1753 574123 • **F:** +44 (0)1753 691632 • enquiries@nfer.ac.uk

www.nfer.ac.uk

NFER ref. DOTS

ISBN. 978-1-912596-04-1