NFER Classroom

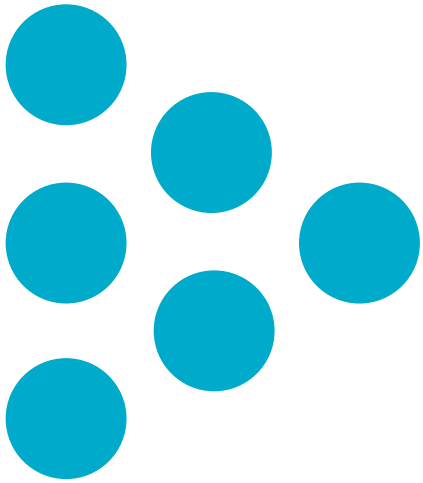
Technical Report

Technical information for NFER Tests in spelling

Suite 2

Centre for Assessment

National Foundation for Educational Research (NFER)



Technical information for NFER tests in spelling Suite 2

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: 978-1-911039-23-5

How to cite this publication:

National Foundation for Educational Research (2020). *Technical Information for NFER Tests in Spelling Suite 2.* Slough: NFER.

Contents

1 Introduction

In January 2015, NFER released a set of spelling assessments and an accompanying teacher guide for Year 5. This was followed by a set for each of Year 3 and Year 4 in January 2016. The finalised materials consist of three standardised spelling tests of equivalent difficulty for each year group. The tests are aligned to the 2014 National Curriculum and are meant to be administered by teachers at three points in the academic year for the purpose of measuring pupil progress. The teacher guide provides information about commonly occurring errors and patterns that occur in misspellings.

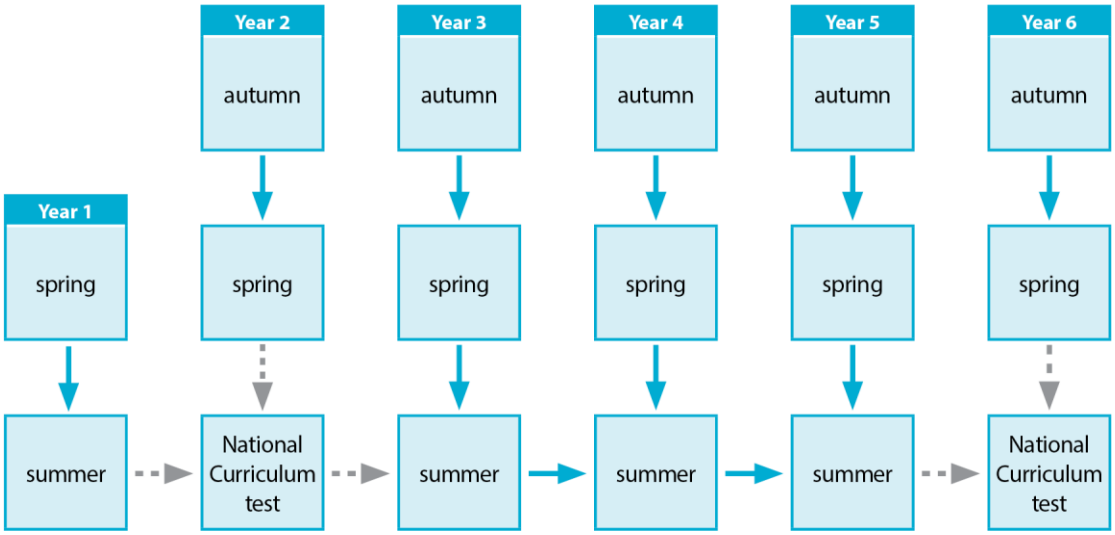
This manual has been published for transparency and to demonstrate the quality of the NFER Tests, so that readers can understand the rigorous development process and appraise the trial data which supports the published materials. It is intended to be of interest to an audience with knowledge of assessment, 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 spelling tests for Years 3 to 5, as these tests were developed in tandem.



3 Early development of items

Following the initial development of word lists, mapped to the National Curriculum, qualitative trialling was conducted at a variety of primary schools. Qualitative trialling involves discussing the 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.

Each spelling test comprises twenty four stand-alone sentences (items) with each sentence testing the spelling of one target word. Target words were selected based on the new (2014) National Curriculum Programme of Study for spelling. For the Year 5 test, the majority of target words were selected based on the statutory requirements in the Year 5 and 6 Programme of Study (PoS), with some coverage of the PoS for spelling in years below Year 5 and some extension into spelling coverage beyond year 6. Similarly, for the Year 3 and 4 tests, the majority of target words were selected based on the statutory requirements in the Year 3 and 4 PoS, with some coverage of the PoS for spelling in years below Year 3 and some extension into spelling coverage beyond Year 4.

4 Standardisation sample characteristics

A large scale standardisation trial was conducted in June 2014 for Year 5 and in June 2015 for Years 3 and 4. Around 4300 pupils participated in the trial of the materials.

The standardisation trial has several purposes. Firstly, it provides item level data from which we can discern exactly how each pupil has 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 2013 (average point score)
- Primary school type
- 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]

[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.

were conducted. The achieved sample representations across the above characteristics are shown and compared with the national population in Tables 1 to 3. The gender breakdown of the sample is shown in Table 4. All data relates to the standardisation trials of 2014 and 2015. The samples were representative of the national population at the school level.

Table 1: Representation of the sample at school level - Year 3 spelling

		population		sample	
		Number	%	Number	%
Primary school type	Primary/Combined	11,057	70	24	59
	Junior	939	6	4	10
	Middle	16	<1	0	0
	Independent schools	1,242	8	3	7
	Other type	2,522	16	10	24
Region	North	4,906	31	11	27
	Midlands	4,810	30	14	34
	South	6,060	38	16	39
KS2 overall performance band 2013 (av. point score)	Lowest 20%	2,560	16	5	12
	2nd lowest 20%	2,389	15	6	15
	Middle 20%	2,616	17	8	20
	2nd highest 20%	2,440	15	11	27
	Highest 20%	2,954	19	7	17
	Missing	2,817	18	4	10
Total schools		**15,776**	**100**	**41**	**100**

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

The Year 3 spelling 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 4 spelling

		population		sample	
		Number	%	Number	%
Primary school type	Primary/Combined	10,916	70	29	62
	Junior	941	6	2	4
	Middle	27	<1	0	0
	Independent schools	1,255	8	4	9
	Other type	2,479	16	12	26
Region	North	4,895	31	14	30
	Midlands	4,711	30	15	32
	South	6,012	38	18	38
KS2 overall performance band 2013 (av. point score)	Lowest 20%	2,531	16	4	9
	2nd lowest 20%	2,369	15	8	17
	Middle 20%	2,605	17	6	13
	2nd highest 20%	2,420	15	10	21
	Highest 20%	2,946	19	15	32
	Missing	2,747	18	4	9
Total schools		**15,618**	**100**	**47**	**100**

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

The Year 4 spelling 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 - Year 5 spelling

		population		sample	
		Number	%	Number	%
Primary school type	Infant/First	12	<1	0	0
	Primary/Combined	11,247	77	37	90
	Junior	987	7	0	0
	Middle	183	1	1	2
	Other type	2,189	15	1	2
	Missing	0	0	2	5
Region	North	4,750	32	15	37
	Midlands	4,522	31	6	15
	South	5,346	37	18	44
	Missing	0	0	2	5
KS2 overall performance band 2013 (av. point score)	Lowest 20%	2,622	18	10	24
	2nd lowest 20%	2,430	17	7	17
	Middle 20%	2,661	18	8	20
	2nd highest 20%	2,442	17	5	12
	Highest 20%	2,954	20	8	20
	Missing	1,509	10	3	7
Total schools		**14,618**	**100**	**41**	**100**

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

The Year 5 spelling 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 4: Representation of the sample at pupil level: gender

	population	Y3 sample		Y4 sample		Y5 sample	
	%	Number	%	Number	%	Number	%
Boys	51	742	51	766	51	688	49
Girls	49	711	49	748	49	688	49
Missing	0	0	0	0	0	17	1
Total	**100**	**1453**	**100**	**1514**	**100**	**1393**	**100**

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

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. Item level data is used primarily to decide which items should comprise the final test. 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 5: Whole test functioning for Year 3, trial versions

	Trial version							
	1	2	3	4	5	6	7	8
Sample *n*	181	226	188	184	159	130	168	217
Reliability (Cronbach's alpha)	0.936	0.927	0.912	0.923	0.910	0.935	0.916	0.919
Maximum score	24	24	24	24	24	24	24	24
Mean	11.52	11.25	12.18	12.28	10.76	10.52	9.46	10.79
Standard deviation	7.16	6.50	6.10	6.71	6.23	6.92	6.18	6.55

Table 6: Whole test functioning for Year 4, trial versions

	Trial version							
	1	2	3	4	5	6	7	8
Sample *n*	204	236	174	183	209	210	216	82
Reliability (Cronbach's alpha)	0.902	0.907	0.894	0.902	0.890	0.898	0.918	0.902
Maximum score	24	24	24	24	24	24	24	24
Mean	10.91	8.58	7.46	8.49	8.72	8.04	8.68	8.40
Standard deviation	6.23	6.03	5.38	5.85	5.48	5.64	6.20	5.71

Table 7: Whole test functioning for Year 5, trial versions

	Trial version							
	1	2	3	4	5	6	7	8
Sample *n*	135	178	186	176	169	183	182	184
Reliability (Cronbach's alpha)	0.920	0.907	0.916	0.900	0.896	0.884	0.907	0.883
Maximum score	24	24	24	24	24	24	24	24
Mean	8.81	9.66	14.36	10.45	10.71	10.91	8.16	8.83
Standard deviation	6.37	6.01	6.16	5.73	5.70	5.36	5.71	5.33

The following tables provide information on the overall functioning of the final tests.

Table 8: Whole test functioning for Year 3, final tests

	Test version		
	1	2	3
Sample n*	1453	1453	1453
Reliability (KR21)	0.923	0.920	0.913
Maximum score	24	24	24
Mean	11.86	12.07	12.05
Standard deviation	7.22	7.11	6.93

Table 9: Whole test functioning for Year 4, final tests

	Test version		
	1	2	3
Sample n*	1514	1514	1514
Reliability (KR21)	0.907	0.913	0.895
Maximum score	24	24	24
Mean	10.02	10.82	10.71
Standard deviation	6.68	6.89	6.46

Table 10: Whole test functioning for Year 5, final tests

	Test version		
	1	2	3
Sample n*	1393	1393	1393
Reliability (KR21)	0.881	0.906	0.880
Maximum score	24	24	24
Mean	11.10	11.60	10.83
Standard deviation	6.20	6.76	6.16

**the total number of pupils in the trial sample for this year group*

6 Item level functioning

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.

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 for gender. 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 is generally balanced.

Table 11 Differential item functioning by gender in the spelling tests

Year	Total number of items	Number of items with no statistically significant DIF	Number of items with a DIF greater than negligible
3	72	64	Girls: 2 Boys: 6
4	72	55	Girls: 9 Boys: 8
5	72	59	Girls: 8 Boys: 5

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. SPOT/SPAT

ISBN. 978-1-911039-23-5