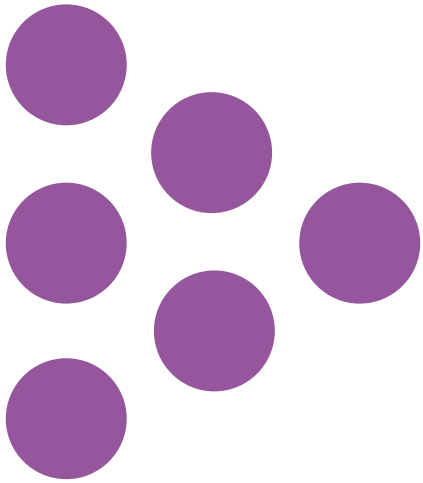
NFER
National Foundation for Educational Research

Report

National Reference Test Results Digest 2018

National Foundation for Educational Research (NFER)



National Reference Test Results Digest 2018

Chris Whetton
Angela Hopkins
Louise Benson

Published in December 2018
By the National Foundation for Educational Research,
The Mere, Upton Park, Slough, Berkshire SL1 2DQ
www.nfer.ac.uk

ISBN: 978-1-911039-85-3

How to cite this publication:

Whetton, C, Hopkins, A and Benson, L. (2018) *National Reference Test Results Digest 2018.* Slough: NFER

NFER
National Foundation for Educational Research

Contents

NFER
National Foundation for Educational Research

1 Introduction

Ofqual has contracted the National Foundation for Educational Research (NFER) to develop, administer and analyse the National Reference Test in English and maths. The first National Reference Test took place in 2017 and established a baseline from which any future changes in standards can be detected. This report represents an overview of the findings of the 2018 testing process.

The National Reference Test (NRT), which consists of a series of test booklets, provides evidence on changes in the performance standards of the same content that is tested in GCSE English language and maths in England at the end of key stage 4. It has been designed to provide additional information to support the awarding of GCSEs in English language and maths and is based on a robust and representative sample of Year 11 students who will, in the relevant year, take their GCSEs.

More information about the NRT can be found in the NRT document collection https://www.gov.uk/government/collections/national-reference-test-information

The first live NRT took place in late February and early March 2017. The outcomes of the 2017 NRT have been benchmarked against the GCSE examinations for that year and a baseline has been established for future years' tests.

The National Reference Test structure is intended to remain the same each year. For each of English and maths there are eight tests in use. All questions are used in two tests, so that effectively all the tests can be analysed together to give a single measure of subject performance. This is similar to other studies that analyse trends in performance nationally, for example, international surveys such as PISA and TIMSS.

This report provides summarised information of the key performance outcomes for English and maths and provides information on the changes from the baseline standards established in 2017. It also includes data on the achievement of the samples, their representativeness and the performance of the students on the tests. Further information on the nature of the tests, the development process, the survey design and its conduct, and the analysis methods used is provided in the accompanying document: ***Background Report: National Reference Test Information***.

2 The Sample

The NRT testing window was 19 February to 2 March 2018. This period was extended by a further week to 9 March, as a result of severe weather conditions which meant that a number of schools were closed during the second week of testing. The rescheduling of test administrations resulted in a small decline in the numbers of schools and students completing the tests in 2018 compared with 2017. The reduction in the number of participating schools may have contributed to a slight reduction in the levels of precision (see sections 4 and 5) but there is no definitive evidence to confirm this. The numbers of schools and students are shown in Table 2.1.

Table 2.1 Target sample sizes and achieved samples in current and previous year

	Target Sample[1]	Achieved Sample	
		Current Year 2018	Previous Year 2017
English			
Number of Schools	330	312	339
Number of Students	7920	6193	7082
Maths			
Number of Schools	330	307	340
Number of Students	7920	6169	7144

The sample was stratified by the achievement level of schools and also school size. In addition, the types of school were monitored. Checks were made on all three of these variables to ensure that the achieved sample was close to that drawn in the sampling frame. This was generally the case, but there was an under-representation of independent schools, whose participation is voluntary. This may have resulted in the final sample of schools being slightly lower attaining than the national population. Given that the sample for the NRT will be drawn on the same basis every year, this arrangement will remain constant each year so it will not impact on the usability of the results.

Table 2.2 shows the number of students in the final sample for whom booklets were dispatched and the number completing the tests for both English and maths. As this shows, around 84 per cent of students took part in the tests. This was a high participation rate and consistent with the rate achieved in 2017.

[1] The total sample of students recruited to participate in the NRT each year is greater than the sample who take the NRT. This is because about a fifth of the students taking the tests are given a booklet which contains a proportion of 'refresh' questions. This is intended to provide contingency questions if needed for future tests. These students' test responses are not included in the sample figures reported in Table 2.1.

Table 2.2 Completed student test returns for English and maths 2018

Test type	No. of students: dispatched tests	No. of students: completed tests	% of students: completed tests
English	7354	6193	84
Maths	7320	6169	84

In total 1,161 students from 292 schools were recorded as non-attendees during the English NRT, which is 16 per cent of the total number of 7,354 sampled students spread across 94 per cent of the schools participating in the survey. Similarly, 1,151 students from 284 schools were recorded as non-attendees during the maths NRT, which is 16 per cent of the total number of 7,320 sampled students spread across 93 per cent of the schools participating in the test. The principal reason given for non-attendance was absence due to illness or other authorised reason, which covered about 60 per cent of non-attendance. Around 10 per cent were in school but did not attend the testing; about five per cent were withdrawn by the headteacher and another five percent had left the school.

The percentage of non-attendance in 2018 was slightly higher than in 2017 when it was 12 per cent. A high student participation rate is needed to ensure precision of the estimates of the results. However, an 84 per cent attendance rate is considered high and there was no evidence that the pattern of non-attendance in 2018 was skewed to particular school types, for example those with lower prior attainment.

The NRT offers access arrangements consistent with JCQ requirements (for GCSE examinations) in order to make the test accessible to as many sampled students as possible. Schools were asked to contact NFER in advance of the NRT to indicate whether any of their students required modified test materials or if students' normal working practice was to use a word processor or laptop during examinations. In cases where additional time would be needed for particular students, schools were asked to discuss this need with the NFER test administrator and ensure that the extra time for the testing session could be accommodated. All requests from schools for access arrangements and the type of arrangement required were recorded. Table 2.3 below shows the different types of access arrangements that were provided to students for the NRT in 2018.

Table 2.3 Number of access arrangements provided 2018

Arrangement provided	No. of students		
	English	Maths	Total
Word processor	297	86	383
Different colour test paper	97	107	204
Modified enlarged print	12	13	25
Enlarged copies	3	5	8
Braille	0	0	0
Total	409	211	620

Note: Table includes the students from the schools that were unable to take to the test due to severe weather conditions. It also includes students in the larger sample who were allocated refresh booklets.

3 Results for the test booklets in 2018

Details of the analysis procedures are given in the accompanying document: ***Background Report: National Reference Test Information***. The analysis process followed a sequence of steps. Initially the tests were analysed using Classical Test Theory to establish that they had performed well, with appropriate difficulty and good levels of reliability. The subsequent analyses used Item Response Theory techniques to link all the tests together and estimate the ability of all the students on a common scale for each subject, independent of the test or items they had taken. These ability estimates were then used for calculating the ability level at the percentiles associated with the GCSE grade boundaries in 2017. From 2018 onwards, the percentages of students achieving above these baseline ability levels are established from the survey.

English

The results of the Classical Test Theory analyses are summarised in Table 3.1. This shows the main test performance statistics averaged for the eight English tests used.

Table 3.1: Summarised Classical Test Theory Statistics for the English Tests in 2018

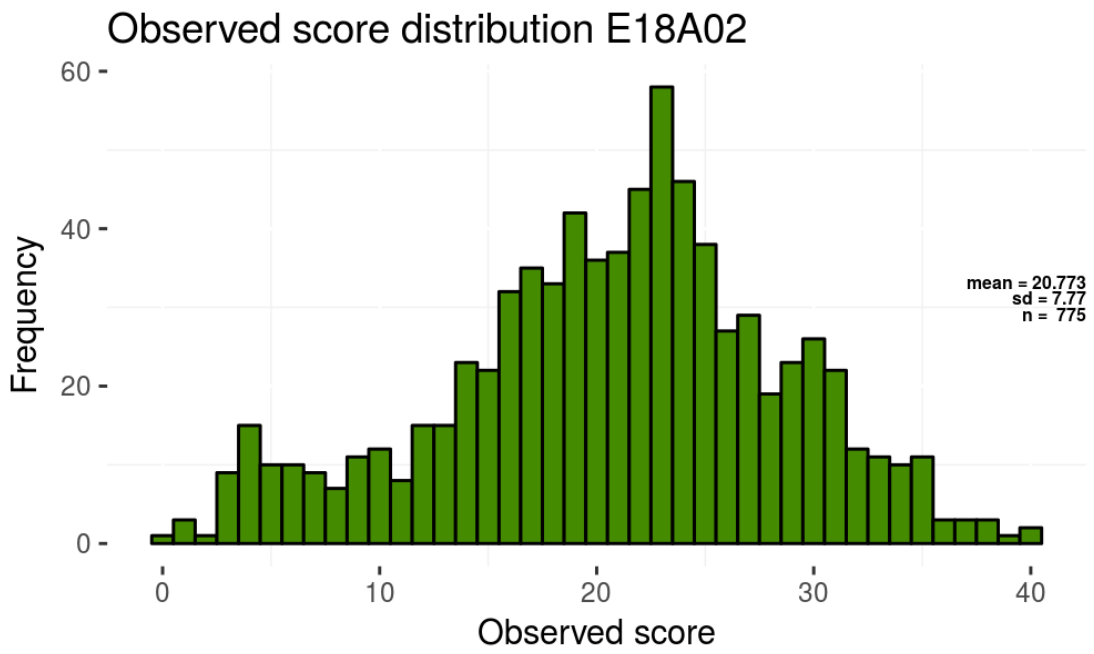
Average Number of Students Taking Each Test	774
Average Maximum Score Attained (out of 50)	42.6
Average Score Attained	20.3
Average Standard Deviation of the Tests	8.3
Average Reliability of the Tests (Coefficient Alpha)	0.77
Average Percentage of Items Attempted by Students	93%

These results show that the English tests functioned well. The maximum scores attained were near the total (total marks available for the paper/booklet) although few students attained scores over 40. The average scores were somewhat less than half marks. The standard deviation shows that the scores were well spread out, allowing discrimination between the students. This is confirmed by the reliability coefficients which are at a good level for an English test of this length. Finally, the average percentage of items attempted by the students at over 90% indicates that the students were engaging with the test and attempting to answer the majority of questions.

These results were confirmed by the distribution of scores which students achieved on the tests. This is shown for one of the tests in Figure 3.1. The distributions were similar for the other tests. The figure shows that scores were attained over nearly all of the possible marks and that the students were fairly evenly spread over the range. It is an example of one test booklet only. There

was a good spread of scores across the available marks, although no students attained the very highest marks.

Figure 3.1: Score Distribution for one of the English Tests



In addition, a full item analysis was carried out for each test, in which the difficulty of every question and its discrimination were calculated. These indicated that all the questions had functioned either well or, in a small number of cases adequately, and there was no need to remove any items from the analyses. All were retained for the Item Response Theory (IRT) analyses. Additionally, an analysis was conducted to establish if any items had performed markedly differently in 2018 compared with 2017. Where there were such indications, possible reasons for this were sought and if there were clear grounds for explaining the change (e.g. a printing error), the items would not be included in the linkage between years. In 2018, no items had to be removed.

Using the common items, the IRT analyses equated the eight tests. The IRT analyses also used the items common between years[2] to equate the tests over years, allowing ability estimates for students in the two years to be on the same scale. After this had been done, the results showed

[2] The 2018 version of the NRT contained the same items as those used in the 2017 test. However, changes were made to the mark scheme for a small number of items in the English test in 2018. For four of these items the changes were very minor and the items were regarded as the same as in 2017. The other four items required changes which were considered significant enough for them to be regarded as new items.

that the mean ability scores for students were very similar for all the tests, confirming that the random allocation to tests had been successful. The results also showed that the level of difficulty of the eight tests was fairly consistent, with only small differences between them. The IRT analyses also indicated that there were no items which functioned differently for male and female students.

Both the Classical Test Theory results and the Item Response Theory results for the English tests showed that these had functioned well to provide good measures of the ability of students, sufficient for estimating averages for the sample as a whole.

Maths

The results of the Classical Test Theory analyses are summarised in Table 3.2. This shows the main test performance statistics averaged for the eight maths tests used.

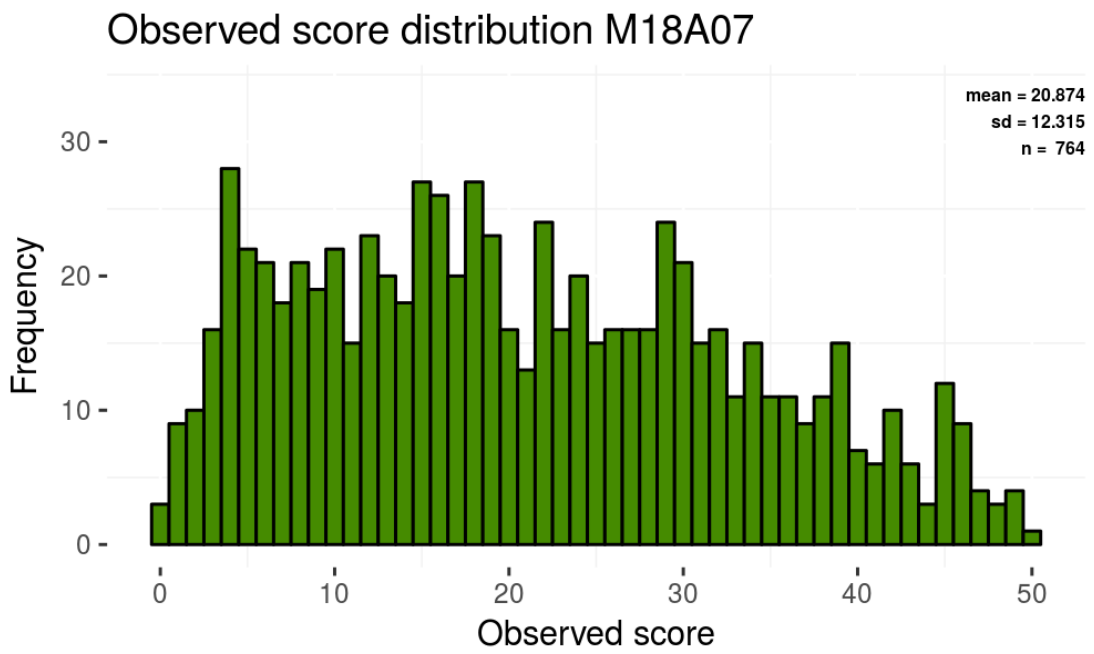
Table 3.2: Summarised Classical Test Theory Statistics for the Maths Tests

Average Number of Students Taking Each Test	771
Average Maximum Score Attained (out of 50)	49.7
Average Score Attained	22.5
Average Standard Deviation of the Tests	12.7
Average Reliability of the Tests (Coefficient Alpha)	0.90
Average Percentage of Items Attempted by Students	88%

These results show that the maths tests also functioned well. The maximum scores attained were just short of the total score possible, and for six tests a small number of students did attain full marks. (This is more likely in maths than English.) The average scores were again slightly less than half marks. The standard deviation shows that the scores were well spread out, allowing discrimination between the students. This is confirmed by the reliability coefficients which are at a good level for a maths test of this length and higher than for English, which again is usual. Finally, the average percentage of items attempted by the students at 88% indicates that the students were engaging with the test and attempting to answer the majority of questions, although to a lesser extent than for the English test. However, there are more items for students to attempt in the maths test.

These results were confirmed by the distribution of scores which students achieved on the tests. This is shown for one of the tests in Figure 3.2. The distributions were similar for the other tests. The figure shows that scores were attained over all of the possible marks and that the students were fairly evenly spread over the range.

Figure 3.2: Score Distribution for one of the Maths Tests



In addition, a full item analysis was carried out for each test, in which the difficulty of every question and its discrimination were calculated. These indicated that all the questions had functioned either well or, in a small number of cases, adequately. There was no need to remove any items from the analyses. All were retained for the Item Response Theory (IRT) analyses. Additionally, an analysis was conducted to establish if any items had performed markedly differently in 2018 compared with 2017. Where there were such indications, possible reasons for this were sought and if there were clear grounds for explaining the change (e.g. a printing error), the items would not be included in the linkage between years. In 2018, no items had to be removed.

Using the common items, the IRT analyses equated the eight tests. The IRT analyses also used the items common between years[3] to equate the tests over years, allowing ability estimates for students in the two years to be on the same scale. After this had been done, the results showed that the mean ability scores for students were very similar for all the tests, confirming that the random allocation to tests had been successful. The results also showed that the level of difficulty

[3] The 2018 version of the NRT contained the same items as those used in the 2017 test. The mark scheme for one maths item was changed but this was very minor.

of the eight tests was fairly consistent, with only small differences between them. The IRT analyses also indicated that there were no items which functioned differently for male and female students.

Both the Classical Test Theory results and the Item Response Theory results for the maths tests showed that these had functioned well to provide good measures of the ability of students, sufficient for estimating averages for the sample as a whole.

Summary

These initial stages of the analyses, the Classical Test Theory evaluation of test functioning and the Item Response Theory equating of the tests indicate that the NRT performed well. This allowed the final stages of the analysis, the estimation of the percentages of students above the same ability thresholds as in 2017 and the calculation of their precision to be undertaken with confidence. These are described in Sections 4 and 5 for English and maths respectively.

4 Performance in English in 2018

The objective of the National Reference Test is to get precise estimates of the percentages of students each year achieving at a level equivalent to three key GCSE grades in 2017: these key grades are 4, 5 and 7. For the NRT in 2017, these baseline percentages were established from the 2017 GCSE population percentages. The NRT ability distribution, based on the IRT analysis, was then used to establish the ability thresholds which corresponded to those percentages. In 2018 and beyond, they will correspond to the same student ability as those of 2017, thus allowing the tracking of standards. Alongside of that, based on the sample achieved and the reliability of the tests, we are able to model the level of precision with which the proportion of students achieving the ability thresholds can be measured. The target for the NRT is to achieve a 95% confidence interval of plus or minus not more than 1.5 percentage points from the estimate at each ability threshold.

Ofqual provided the percentages of students at or above the three relevant grades (grades 4, 5 and 7) taken from the 2017 GCSE population. These are shown in Table 4.1. These percentages were mapped to three ability threshold scores in the NRT in 2017.

Table 4.1 English 2017 NRT Baseline Thresholds used for the 2018 NRT

	Percentage of students above threshold from 2017 GCSE
Grade 7 and above	16.8
Grade 5 and above	53.3
Grade 4 and above	69.9

For 2018, the 2017 and 2018 NRT data were analysed together using Item Response Theory (IRT) modelling techniques. By analysing all the data concurrently, ability distributions could be produced for the 2017 and 2018 samples on the same scale. The percentages of students at each of the three GCSE grade boundaries, fixed on the 2017 distribution, could then be mapped onto the 2018 distribution to produce estimates of the percentage of students at the same level of ability in the 2018 sample. For example, the percentage of pupils at the 'Grade 4 and above' threshold in the 2017 GCSE population was 69.9%. This was mapped onto the 2017 distribution to read off an ability value equivalent to that grade boundary (in this case, -0.943, although the value itself is not meaningful in its own right). The same ability value (-0.943) on the 2018 distribution can then be found, and the percentage of students at this threshold or above in the 2018 sample read off (68.5%). In this way, we are able to estimate the percentage of students at the same level of ability as represented in the 2017 GCSE population, for each year of the NRT going forward. The precision of these estimates is dependent on both the sample achieved and the reliability of the tests as measures.

Table 4.2 and Figures 4.1 and 4.2 summarise the outcomes of the 2018 English NRT at the three key grade boundaries.

Table 4.2 shows the percentage of students achieving above the NRT ability level associated with each of the three key grades. Confidence intervals for percentages of students above the ability thresholds are provided in brackets alongside the estimates for 2018. The half-width of the confidence interval for the percentages is also provided in a separate column, to give an overall indicator of the level of precision of the outcomes. This is important as it shows that although there has been a small decline in performance for the two lower grade boundaries and a very slight increase in the grade 7 boundary, the 2018 confidence intervals for all three grades include the 2017 baseline percentages. This means that the change cannot be regarded as statistically significant.

Table 4.2 NRT Percentage of students reaching threshold in 2017 and 2018

Threshold	Percentage of students above threshold in 2017	Estimated percentage of students above threshold in 2018 (with 95% confidence interval)	Half-width of 95% confidence interval in percentage points
Grade 7 and above	16.8	16.9 (15.5 to 18.3)	±1.4
Grade 5 and above	53.3	52.6 (50.5 to 54.8)	±2.2
Grade 4 and above	69.9	68.5 (66.6 to 70.4)	±1.9

The data are shown graphically in Figure 4.1, which shows that the percentage above each grade decreased slightly in 2018 for the two lower grade boundaries and increased very slightly for the grade 7 boundary. These changes were not statistically significant.

Figure 4.1 NRT percentage of students reaching threshold in 2017 and 2018, including the 95% confidence interval

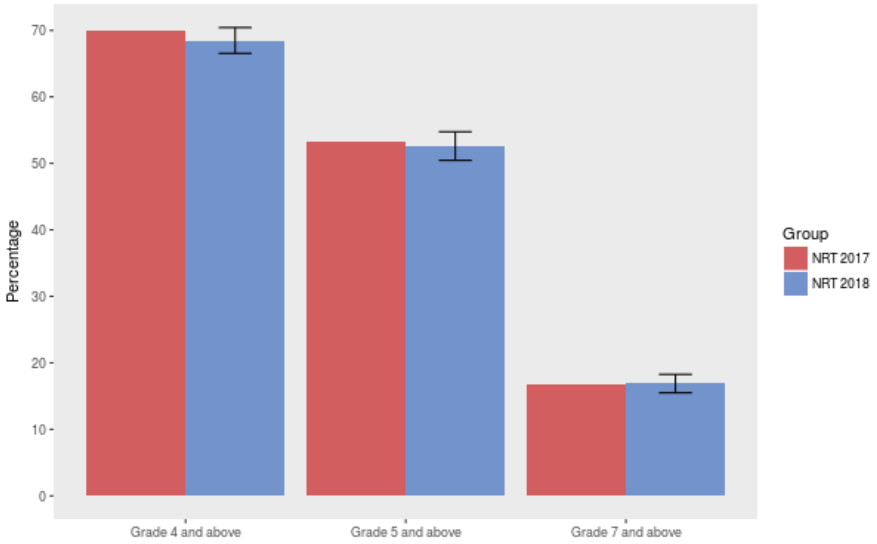
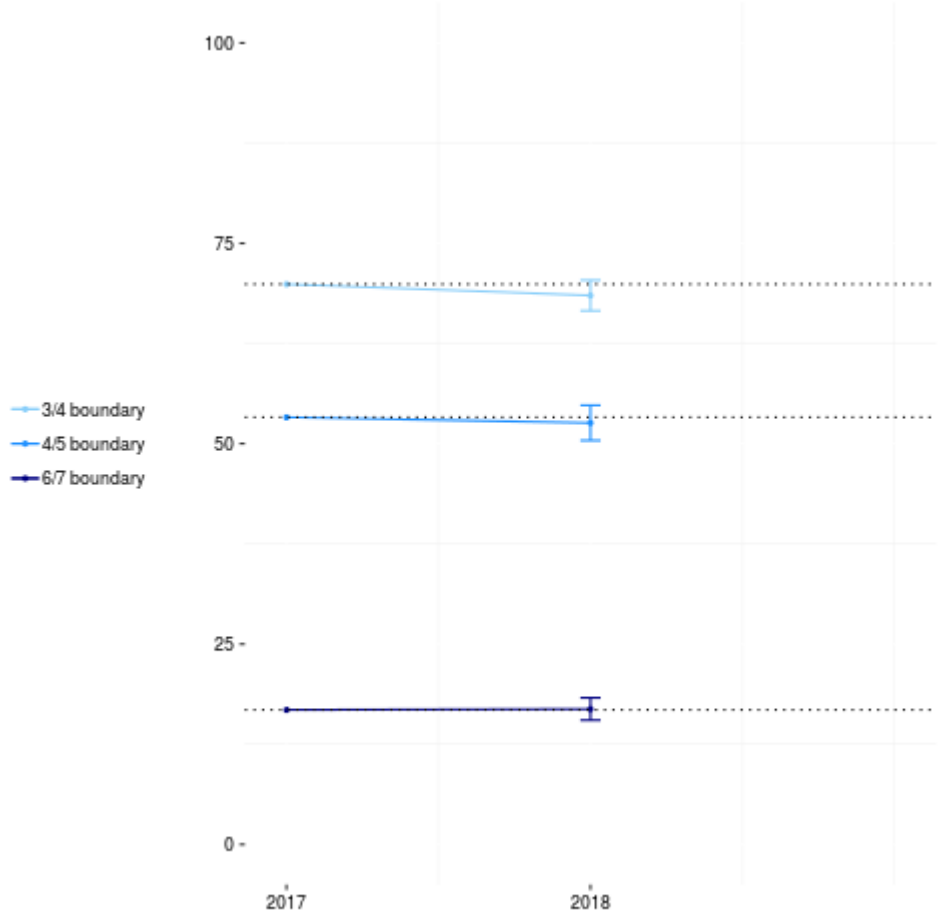


Figure 4.2 presents 95% confidence intervals around the percentages achieving at least the specified grade boundary in 2018, as compared to the 2017 population baseline percentages. The 2017 population percentages are represented as dotted lines and the trend lines across years as solid lines. This format has been used to encourage the reader to compare the 2018 point estimate confidence bands with the 2017 baseline population percentages. Again this shows that the slight changes in percentages cannot be considered as significant. The format will become more useful as future years are added, giving a long-term time series.

Figure 4.2 Long term changes in NRT English over time from 2017 baseline



5 Performance in maths in 2018

The objective of the National Reference Test is to get precise estimates of the percentages of students each year achieving at a level equivalent to three key GCSE grades in 2017: these key grades are 4, 5 and 7. For the NRT in 2017, these baseline percentages were established from the 2017 GCSE population percentages. The NRT ability distribution, based on the IRT analysis, was then used to establish the ability scores which corresponded to those percentages. In 2018 and beyond, they will correspond to the same student ability as those of 2017, thus allowing the tracking of standards. Alongside of that, based on the sample achieved and the reliability of the tests, we are able to model the level of precision with which the proportion of students achieving the ability scores can be measured. In this context, the precision is half of the 95 per cent confidence intervals for the measurement of these percentages. The target for the NRT is to achieve a 95% confidence interval of plus or minus not more than 1.5 percentage points from the estimate at each ability threshold.

Ofqual provided the percentages of students at or above three relevant grades (grades 4, 5 and 7) taken from the 2017 population. These are shown in Table 5.1. These percentages were mapped to three ability threshold scores in the NRT in 2017.

Table 5.1 Maths 2017 NRT Baseline Thresholds used for the 2018 NRT

	Percentage of students above threshold from 2017 GCSE
Grade 7 and above	19.9
Grade 5 and above	49.7
Grade 4 and above	70.7

For 2018, the 2017 and 2018 NRT data were analysed together using Item Response Theory (IRT) modelling techniques. By analysing all the data concurrently, ability distributions could be produced for the 2017 and 2018 samples on the same scale. The percentages of students at each of the three GCSE grade boundaries, fixed on the 2017 distribution, could then be mapped onto the 2018 distribution to produce estimates of the percentage of students at the same level of ability in the 2018 sample. For example, the percentage of pupils at the 'Grade 4 and above' threshold in the 2017 GCSE population was 70.7%. This was mapped onto the 2017 distribution to read off an ability value equivalent to that grade boundary (in this case, -0.846, although the value itself is not meaningful in its own right). The same ability value (-0.846) on the 2018 distribution can then be found, and the percentage of students at this threshold or above in the 2018 sample read off (73.1%). In this way, we are able to estimate the percentage of students at the same level of ability as represented in the 2017 GCSE population, for each year of the NRT going forward. The

precision of these estimates is dependent on both the sample achieved and the reliability of the tests as measures.

Table 5.2 and Figures 5.1 and 5.2 summarise the outcomes of the 2018 maths NRT at the three key grade boundaries.

Table 5.2 shows the percentage of students achieving above the NRT ability level associated with each of the three key grades. This shows that the percentage of students above each grade increased in 2018. Confidence intervals for percentages of students above the ability thresholds are provided in brackets alongside the estimates for 2018. The half-width of the confidence interval for the percentages is also provided in a separate column, to give an overall indicator of the level of precision of the outcomes. The lower bounds of the 2018 confidence intervals for all three grades lie above the 2017 baseline percentages. This means that the increase can be regarded as statistically significant at all three grade boundaries.

Table 5.2 Maths NRT Percentage of students reaching threshold in 2017 and 2018

Threshold	Percentage of students above threshold in 2017	Estimated percentage of students above threshold in 2018 (with 95% confidence interval)	Half-width of 95% confidence interval in percentage points
Grade 7 and above	19.9	21.9 (20.4 to 23.4)	±1.5
Grade 5 and above	49.7	52.6 (50.8 to 54.3)	±1.7
Grade 4 and above	70.7	73.1 (71.6 to 74.6)	±1.5

The data are shown graphically in Figure 5.1, which shows that the percentage above each grade increased in 2018 and that this was statistically significant.

Figure 5.1 Percentage of student abilities that exceeded the relevant grades in maths, including the 95% confidence interval

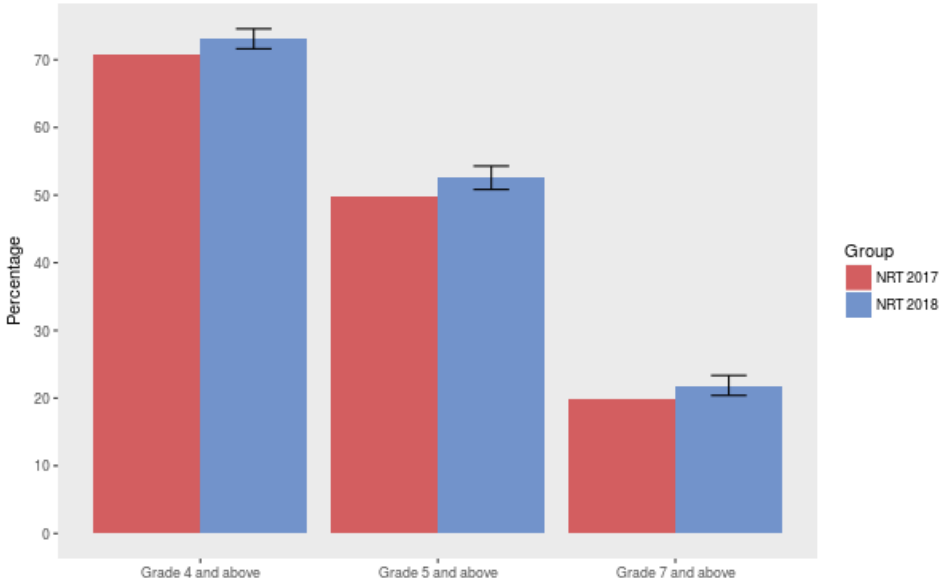
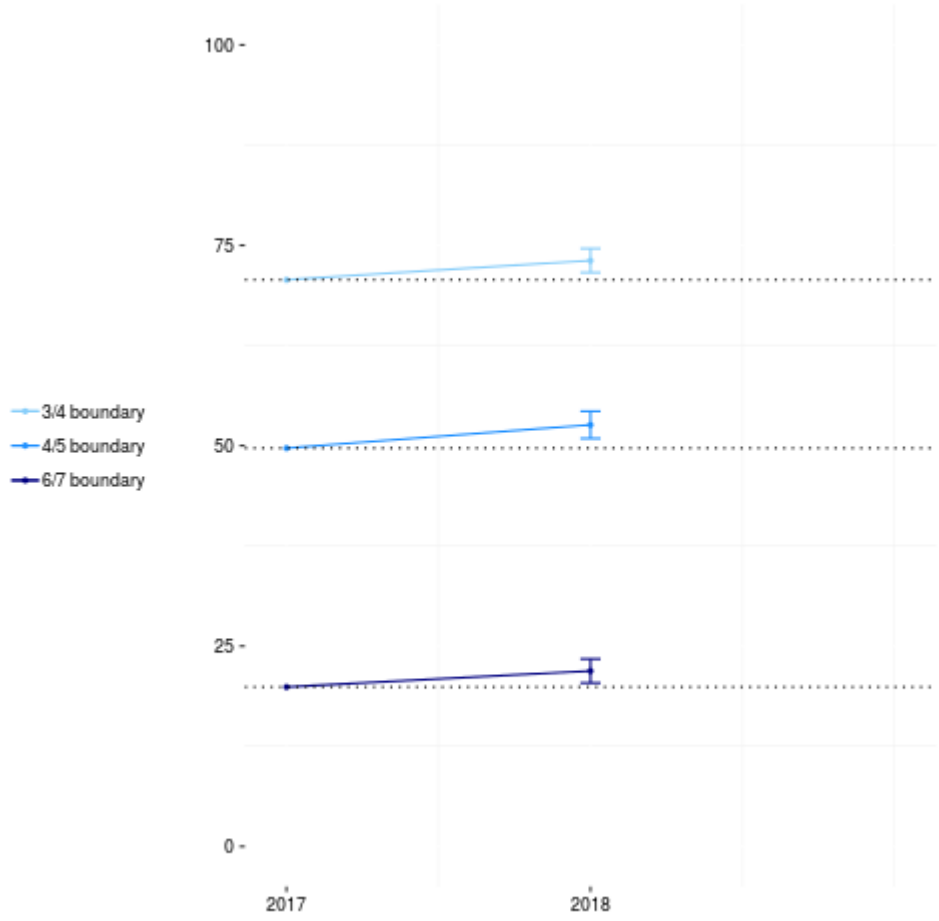


Figure 5.2 presents the same data in a different diagrammatic format. It shows the 95% confidence intervals around the percentages achieving at least the specified grade boundary in 2018, as compared to the 2017 population baseline percentages. The 2017 population percentages are represented as dotted lines and the trend lines across years as solid lines. This format has been used to encourage the reader to compare the 2018 point estimate confidence bands with the 2017 baseline population percentages. Again, this shows that the increase in percentages can be considered as statistically significant at all three boundaries of interest. The format will become more useful as future years are added, giving a long-term time series.

Figure 5.2 Long term changes in NRT Maths over time from 2017 baseline



6 Appendix A: a brief summary of the NRT

English

The English test takes one hour to administer and follows the curriculum for the reformed GCSE in English language. In each of the eight English test booklets, there are two components; the first is a reading test and the second a writing test. Each component carries 25 marks and students are advised to spend broadly equal time on each component.

The reading test is based on an extract from a longer prose text, or two shorter extracts from different texts. Students are asked five, six or seven questions that refer to the extracts. Some questions of one to four marks require short responses or require the student to select a response from options provided. In each booklet, the reading test also includes a 6-mark question and a 10-mark question where longer, more in-depth responses need to be given. These focus on analysis and evaluation of particular aspects of the text.

The writing test is a single, 25-mark task. This is an extended piece of writing, responding to a stimulus. For example, students may be asked to describe, narrate, give and respond to information, argue, explain or instruct.

Maths

For maths, a separate sample of students is also given one hour to complete the test. The test includes questions on number, algebra, geometry and measures, ratio and proportion, and statistics and probability – the same curriculum as the reformed GCSE. Each of the eight test booklets has 13 or 14 questions with a total of 50 marks and each student takes just one of the test booklets.

Analysis

The analysis process followed a sequence of steps. Initially, the tests were analysed using Classical Test Theory to establish that they had performed well, with appropriate difficulty and good levels of reliability. The subsequent analyses used Item Response Theory techniques to link all the tests together across 2017 and 2018 and estimate the ability of all the students on a common scale for each subject for each year, independent of the test or items they had taken. These ability estimates were then used for calculating the ability level at the percentiles associated with the GCSE grade boundaries in 2017 and mapping these onto 2018 to generate percentile estimates for 2018.

National Foundation for Educational Research

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. OFMT