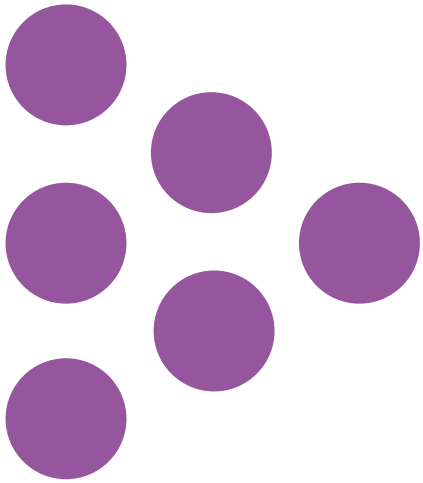
NFER
National Foundation for Educational Research

Report

National Reference Test Results Digest 2020

National Foundation for Educational Research (NFER)



National Reference Test Results Digest 2020

Bethan Burge
Louise Benson

Published in August 2020

By the National Foundation for Educational Research,

The Mere, Upton Park, Slough, Berkshire SL1 2DQ

www.nfer.ac.uk

ISBN: 978-1-911039-90-7

How to cite this publication:

Burge, B and Benson, L. (2020) *National Reference Test Results Digest 2020.* 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 (NRT) in English and maths. The first NRT 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 2020 testing process.

The NRT, which consists of a series of test booklets, provides evidence on changes in the performance standards in GCSE English language and maths in England at the end of key stage 4. It does this by testing content taken from the revised GCSE English and maths curricula. 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[1].

More information about the NRT can be found in the NRT document collection.

The first live NRT took place in late February and early March 2017. The outcomes of the 2017 GCSE examinations for that year provided the baseline percentages of students at three grade boundaries and these were mapped to the NRT for 2017 to establish the corresponding proficiency level. The percentages of students achieving those proficiency levels in each subsequent year are calculated and compared.

The National Reference Test structure is intended to remain the same each year. For each of English and maths there are eight test booklets in use. Each question is used in two booklets, 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 over time, for example, international surveys such as PISA and TIMSS.

This report provides summarised information of the key performance outcomes for English and maths in 2020 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***.

[1] In 2020 the Year 11 cohort did not take their GCSE examinations due to the school closures that occurred in a bid to fight the spread of Covid-19. Instead GCSE grades will be awarded based on an assessment of the grade these students would have been most likely to achieve had examinations gone ahead and following a statistical standardisation process. The government's intention is that results would be issued to this year's cohort based on a range of evidence and data, including performance on mock examinations and non-examination assessment. (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/890811/Summer_2020_grades_for_GCSE_AS_A_level_guidance_for_teachers_students_parents_09062020.pdf)

2 The sample

The NRT took place between 24 February and 6 March 2020. The numbers of participating schools and students are shown in Table 2.1. The number of schools in the sample was very similar to 2019, and above target, but the number of students was slightly lower.

Table 2.1. Target sample sizes and achieved samples in current and previous years

	Target Sample	Achieved Sample			
		Current Year 2020	Previous Year 2019	Previous Year 2018	Previous Year 2017
English					
Number of Schools	330	332	332	312	339
Number of Students	6732*	6639	6740	6193	7082
Maths					
Number of Schools	330	333	331	307	340
Number of Students	6732*	6756	6826	6169	7144

*The target number of students is based on an attendance rate of 85 per cent of the total number of students if the maximum number of schools are recruited.

The sample was stratified by the previous attainment of schools in GCSE English language and maths and also by 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 in the achieved sample, probably because their participation is voluntary. There was also a slight over-representation of academies in 2020. The under-representation of independent schools may have resulted in the final sample being slightly lower attaining than the national population but this was also true in previous years. The slight difference between the achieved sample and the sample frame at the top end of the distribution based on previous school GCSE performance has remained broadly consistent across the years.

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 85 per cent of students who were selected took part in the tests. This was a high participation rate and consistent with the rates achieved in 2017, 2018 and 2019.

Table 2.2. Completed student test returns for English and maths 2020

Test type	No. of students: dispatched tests*	No. of students: completed tests	% of students: completed tests
English	7845	6639	85
Maths	7886	6756	86

*This is lower than the number of students sampled as the student sample for both English and Maths includes a school that withdrew from testing due to snow and one pupil from another school who was withdrawn from testing prior to dispatch.

In total 1,206 students from 320 schools were recorded as non-attendees during the English NRT, which is 15 per cent of the total number of 7,870 sampled students spread across 96 per cent of the schools participating in the assessment. For maths, a total of 1,130 students from 306 schools were recorded as non-attendees during the maths NRT, which is 14 per cent of the total number of 7,908 sampled students spread across 92 per cent of the schools participating in the test.

The pattern of non-attendance is similar in maths to English. The principal reason given for non-attendance was absence due to illness or other authorised reason, which covered 58 per cent of non-attendance for English and 55 per cent of non-attendance for maths. Absent from the testing session but present in school remains the second most frequently recorded reason, 12 per cent for English and 15 per cent for maths. Of the remaining reasons for non-attendance, about nine per cent of students were withdrawn by the headteacher and another six per cent had left the school.

The percentage of non-attendance in 2020 was similar to that in 2019. A high student participation rate is needed to ensure precision of the estimates of the results. However, an 85 per cent attendance rate is considered high when compared to other monitoring tests such as international large-scale assessments and there was no evidence that the pattern of non-attendance was skewed to particular school types, for example, those with lower performance in previous English language and maths GCSEs.

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 arrangement that were provided to students for the NRT in 2020. These are the access arrangements facilitated by NFER for the NRT in 2020, we do not collect complete data on the permitted arrangements which are made by schools. Overall, the number of access arrangements provided were similar to 2019. However, there were fewer laptops and online tests provided in 2020 in comparison to previous NRTs and the number of different coloured test papers increased.

Table 2.3. Number of access arrangements provided 2020

Arrangement provided	No. of students		
	English	Maths	Total
Online test	209	126	335
NFER laptop	25	9	34
Different colour test paper	101	135	236
Modified enlarged print	13	18	31
Enlarged copies	5	3	8
Braille	0	0	0
Total	353	291	644

NB: Due to some students having multiple access arrangements they will be featured twice in the table

3 Results for the test booklets in 2020

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 (IRT) 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 NRT.

English

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

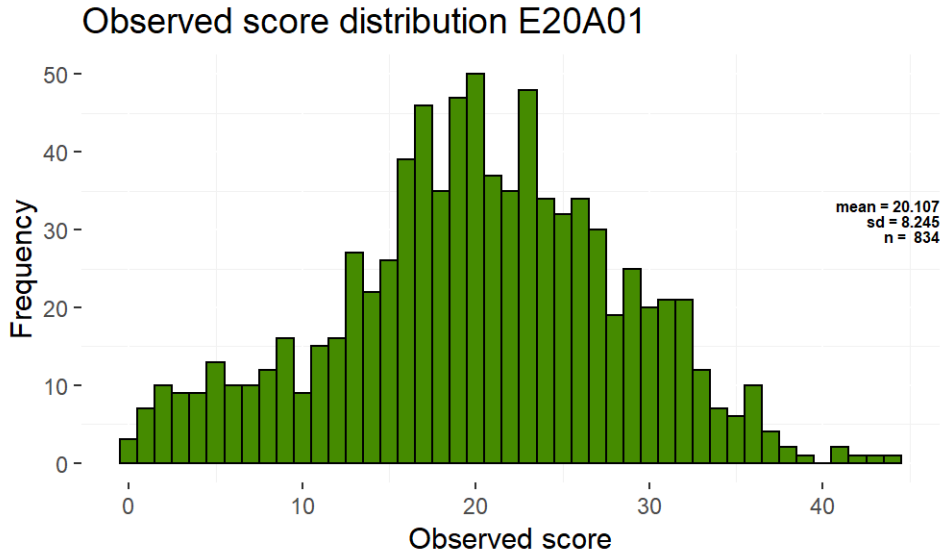
Table 3.1. Range of Classical Test Theory statistics for the English tests in 2020

	Minimum	Maximum
Number of Students Taking Each Test Booklet	814	840
Maximum Score Attained (out of 50)	40	45
Average Score Attained	19.0	20.8
Standard Deviation of the Test Booklets	7.7	9.4
Reliability of the Tests (Coefficient Alpha)	0.73	0.79
Average Percentage of Items Attempted by Students (%)	92	94

These results show that the English test booklets functioned well. The maximum scores attained were near the total marks available for the booklet, although few students attained scores over 40. The average scores were somewhat less than half the available 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 per cent for all booklets 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 students achieved on the tests. This is shown for one of the tests in Figure 3.1. It is an example of one test booklet only but 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 spread across the range, 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. Therefore all were retained for the IRT analyses. Additionally, an analysis was conducted to establish if any items had performed markedly differently in 2020 compared with the previous years. Where there are such indications, a formal procedure is followed for reviewing the items to establish whether there could be an external reason for the change. For 2020, two English items were removed from the link between 2017-19 and 2020.

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 all four 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 of the eight tests was fairly consistent, with only small differences between them.

Both the Classical Test Theory results and the IRT 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.

[2] The 2020 version of the NRT contained the same items as those used in the tests from 2017 to 2019.

Maths

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

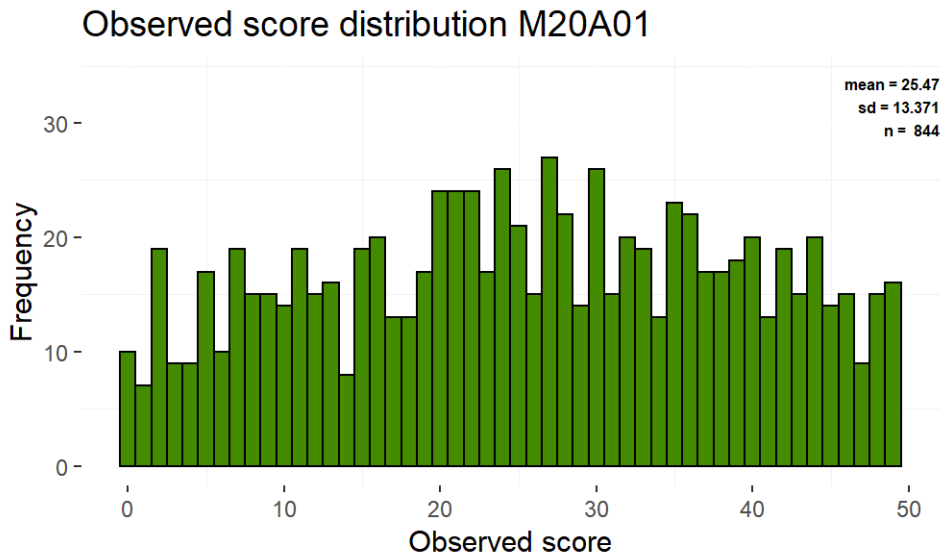
Table 3.2. Range of Classical Test Theory statistics for the maths tests in 2020

	Minimum	Maximum
Number of Students Taking Each Test Booklet	832	866
Maximum Score Attained (out of 50)	49	50
Average Score Attained	21.5	25.5
Standard Deviation of the Test Booklets	12.1	13.8
Reliability of the Tests (Coefficient Alpha)	0.90	0.92
Average Percentage of Items Attempted by Students (%)	85	90

These results show that the maths tests also functioned well. The maximum score, or one mark short of it, was attained on all booklets. The average scores were, again, slightly less than half marks for most booklets. 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 very 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 close to 90 per cent 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 the range of 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 IRT analyses. Additionally, an analysis was conducted to establish if any items had performed markedly differently in 2020 compared with the previous years. Where there are such indications, a formal procedure is followed for reviewing the items to establish whether there could be an external reason for the change and if there is sufficient evidence to remove the item from the link between years. In 2020, one maths item was removed from the link between 2017 and 2018-2020, following new evidence of a change in marking severity at that time.

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 all four years to be on the same scale. After this had been done, the results showed that the mean ability scores for students were 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.

Both the Classical Test Theory results and the IRT 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.

[3] The 2020 version of the NRT contained the same items as those used in the tests from 2017 to 2019.

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 2020

The objective of the National Reference Test (NRT) 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 Item Response Theory (IRT) analysis, was then used to establish the ability thresholds which corresponded to those percentages. From 2018 onwards, the thresholds correspond to the same level of student ability as the thresholds established in 2017, thus allowing us to estimate the percentage of students above each of those thresholds and track performance over time. Alongside this, 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

Threshold	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

In 2020, the NRT data for the years 2017 to 2020 were analysed together using IRT modelling techniques. By analysing all the data concurrently, ability distributions could be produced for the samples for each year 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 distributions for the subsequent years to produce estimates of the percentage of students at the same level of ability in those years. For example, the percentage of students at the 'Grade 4 and above' threshold in the 2017 GCSE population was 69.9 per cent. This was mapped onto the 2017 distribution to read off an ability value at that grade boundary. The same ability value on the 2018, 2019 and 2020 distributions can then be found, and the percentage of students at this threshold or above in those years established. 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 presents the percentages of students achieving above the specified grade boundaries for the years 2017 to 2020. Confidence intervals for percentages are provided in brackets alongside the estimates. This is important as it shows that although there have been changes in

performance, these are often within the confidence intervals. The statistical interpretation of the differences is discussed below.

Table 4.2. Estimated percentages at grade boundaries in English

Year	Grade 4 and above	Grade 5 and above	Grade 7 and above
2017	69.9 (68.0 - 71.8)	53.3 (51.4 - 55.1)	16.8 (15.5 - 18.1)
2018	68.3 (66.1 - 70.6)	52.5 (50.5 - 54.6)	16.8 (15.4 - 18.3)
2019	65.7 (63.9 - 67.4)	49.6 (47.5 - 51.6)	16.2 (14.7 - 17.6)
2020	67.0 (65.3 - 68.8)	51.5 (49.6 - 53.3)	17.5 (16.2 - 18.9)

The 2017 figures in the table above are based on the NRT study, rather than the 2017 GCSE percentages. Note that, because of the way in which they have been computed, they match closely with the GCSE percentages. The confidence intervals for them reflect the fact that the NRT 2017 outcomes carry the statistical error inherent in a sample survey, as per the subsequent years.

Since the percentages for previous years have been re-estimated following the concurrent calibration with the 2020 data, these figures differ slightly from those reported in previous years. Some degree of variation is expected given the addition of more data, and the differences seen are small.

Table 4.3 shows the half widths of the confidence intervals, which in most cases are close to the 1.5 percentage points target. The table illustrates that the precision for 2020 is relatively consistent with that in 2019.

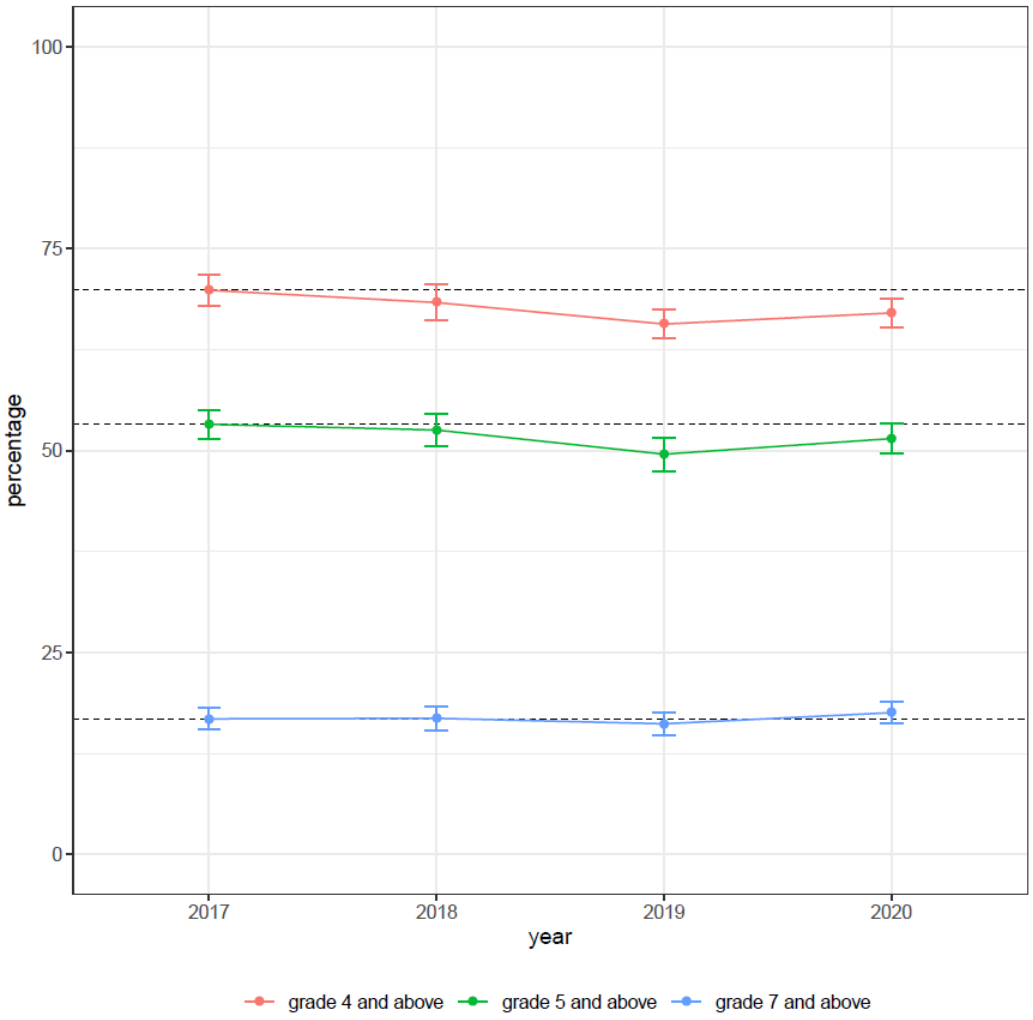
Table 4.3. English NRT half width of confidence intervals each year

Year	Grade 4 and above	Grade 5 and above	Grade 7 and above
2017	1.9	1.8	1.3
2018	2.3	2.0	1.5
2019	1.8	2.1	1.4
2020	1.8	1.8	1.3

Figure 4.1 presents 95% confidence intervals around the percentages achieving at least the specified grade boundary in 2020, as compared to previous years and 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 point estimate confidence bands for each year with the 2017 baseline population percentages, bearing in mind the confidence intervals.

Figure 4.1. Long term changes in NRT English over time from 2017 baseline



The chart shows a decline in performance across the first three years of the NRT, more notably for grade 4 and above and grade 5 and above, followed by an improvement in 2020. A key question arising for the NRT results in a given year is to determine if differences in outcomes across the years are statistically significant. For the NRT, several comparisons could be made between different pairs of years at different grade boundaries, and this gives rise to a danger that changes that arise by chance may seem real. Hence the criteria for significance which have been used are adjusted for multiple comparisons. (For more information see Appendix A.)

The research question NFER was asked to address is to compare the performance in 2020 with the performance in 2017 at each of the three grade boundaries. Adjusting for three comparisons, the NRT English data shows that there are no significant differences in performance between 2017 and 2020 at any of the three grade boundaries.[4]

[4] The results of a given year's NRT can be compared with the NRT study of 2017 (both are sample surveys, and the statistical error is therefore reflected in confidence intervals for 2017) or with the GCSE percentages of 2017, regarded as external constants. The *2018 Results Digest* reported comparisons with the GCSE 2017 population percentages. However, in order to make ongoing comparisons from year to year it was decided for 2019 onwards that comparing the outcomes between NRT studies (i.e. making statistical comparisons with the 2017 NRT study, rather than 2017 GCSE percentages) would be more informative.

5 Performance in maths in 2020

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 Item Response Theory (IRT) analysis, was then used to establish the ability scores which corresponded to those percentages. From 2018 onwards, the thresholds correspond to the same level of student ability as the thresholds established in 2017, thus allowing us to estimate the percentage of students above each of those thresholds and track performance over time. Alongside this, 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. 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 GCSE 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

Threshold	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

In 2020, the NRT data for the years 2017 to 2020 were analysed together using IRT modelling techniques. By analysing all the data concurrently, ability distributions could be produced for the 2017, 2018 and 2019 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 distributions for the subsequent years to produce estimates of the percentage of students at the same level of ability in those years. For example, the percentage of students at the 'Grade 4 and above' threshold in the 2017 GCSE population was 70.7 per cent. This was mapped onto the 2017 distribution to read off an ability value equivalent to that grade boundary. The same ability value on the 2018, 2019 and 2020 distributions can then be found, and the percentage of students at this threshold or above in those years established. 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 presents the percentages of students achieving above the specified grade boundaries for the years 2017 to 2020. Confidence intervals for percentages are provided in brackets alongside

the estimates. This is important as it shows that although there have been changes in performance, these are often within the confidence intervals. The statistical interpretation of the differences is discussed below.

Table 5.2. Estimated percentages at grade boundaries in maths

Year	Grade 4 and above	Grade 5 and above	Grade 7 and above
2017	70.7 (69.4 - 72.1)	49.7 (48.2 - 51.3)	19.9 (18.5 - 21.3)
2018	73.2 (71.8 - 74.6)	52.6 (50.9 - 54.2)	21.5 (20.2 - 22.8)
2019	73.3 (71.9 - 74.6)	52.1 (50.5 - 53.8)	22.9 (21.5 - 24.2)
2020	74.0 (72.6 - 75.4)	54.4 (52.9 - 55.9)	24.0 (22.7 - 25.3)

The 2017 figures in the table above are based on the NRT study, rather than the 2017 GCSE percentages. Note that, because of the way in which they have been computed, they match closely with the GCSE percentages. The confidence intervals for them reflect the fact that the NRT 2017 outcomes carry the statistical error inherent in a sample survey, as per the subsequent years.

Since the percentages for previous years have been re-estimated following the concurrent calibration with the 2020 data, these figures differ slightly from those reported in previous years. Some degree of variation is expected given the addition of more data, and the differences seen are small.

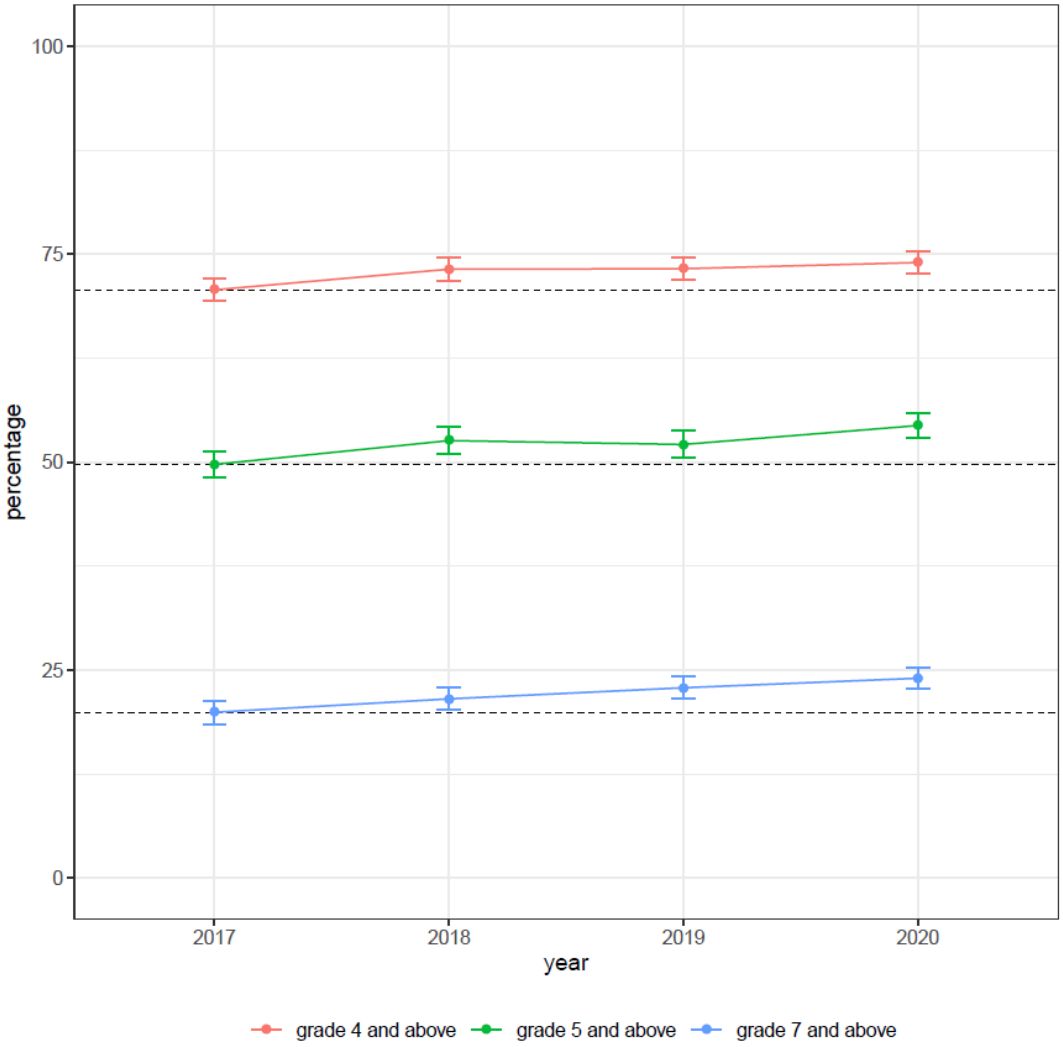
Table 5.3 shows the half widths of the confidence intervals, which in most cases are lower than or close to the 1.5 percentage points target. The table illustrates that the precision for 2020 is relatively consistent with that in 2019.

Table 5.3. Maths NRT half width of confidence intervals each year

Year	Grade 4 and above	Grade 5 and above	Grade 7 and above
2017	1.3	1.6	1.4
2018	1.4	1.7	1.3
2019	1.3	1.6	1.3
2020	1.4	1.5	1.3

Figure 5.1 presents 95% confidence intervals around the percentages achieving at least the specified grade boundary in 2020, as compared to previous years and 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 point estimate confidence bands for each year with the 2017 baseline population percentages, bearing in mind the confidence intervals.

Figure 5.1. Long term changes in NRT maths over time from 2017 baseline



The chart shows a steady increase across all years of the NRT at all three grade boundaries. A key question arising for the NRT results in a given year is to determine if differences in outcomes across the years are statistically significant. For the NRT, several comparisons could be made and this gives rise to a danger that changes that arise by chance may seem real. Hence the criteria for significance which have been used are adjusted for multiple comparisons. (For more information see Appendix A.)

The research question NFER was asked to address is to compare the performance in 2020 with the performance in 2017 at each of the three grade boundaries. Adjusting for three multiple comparisons, the NRT maths data shows that there has been a statistically significant increase in performance between 2017 and 2020 at all three grade boundaries at the 1% level of significance.[5]

[5] The results of a given year's NRT can be compared with the NRT study of 2017 (both are sample surveys, and the statistical error is therefore reflected in confidence intervals for 2017) or with the GCSE percentages of 2017, regarded as external constants. The *2018 Results Digest* reported comparisons with the GCSE 2017 population percentages. However, in order to make ongoing comparisons from year to year it was decided for 2019 onwards that comparing the outcomes between NRT studies (i.e. making statistical comparisons with the 2017 NRT study, rather than 2017 GCSE percentages) would be more informative.

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 extract(s). 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 or a comparison between texts.

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 from 2017 to 2020 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 the distributions for subsequent years to generate percentile estimates for those years.

Multiple Comparisons

The statistical significance of the difference between two percentages estimated in two years, say 2017 and 2020, may be approached with a two-sample t-statistic. Because of the huge number of degrees of freedom, the value can be compared with the standard normal distribution rather than the t-distribution. For a comparison of two percentages, say the percentage of students at grade 4 or higher between two years, the critical value at a confidence level of 0.05 (5%) would usually be 1.96. However, since there are three grade thresholds across multiple years, there are a number of

comparisons which could be made (up to 18 if all pairs of years were compared across all three grade boundaries). As the number of simultaneous comparisons grows, the probability that some of them are significant by chance rapidly increases. To guarantee that the chosen level of significance is guaranteed overall, we have implemented an adjustment for multiple comparisons.

NFER
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