SUMMARY

NRT Annual Statement 2019

Introduction

In February/March 2019 just over 13,500 year 11 students from over 330 schools took the third annual National Reference Test (NRT) in English and maths, which is administered by NFER. The tests are designed to provide evidence on the performance of 16-year-old students in English language and maths. The first live NRT, taken in 2017, was benchmarked against the first awards of the reformed GCSEs in English language and maths, and subsequent tests compare the performance of students with those in 2017.

Results are reported at three grade boundaries – grade 7, grade 5 and grade 4. Results are reported as expected percentages of students achieving those grades (and above) based on changes in performance on the NRT. This report focuses only on grades 7 and 4, since grade 5 is an arithmetic grade and would not normally be adjusted.

Results for 2019

The results are shown below. Because this test uses a sample of students, we report 'confidence intervals' around the results. These confidence intervals represent the possibility that if we had taken a sample of different students, and each student had taken a different subset of questions, we would get a slightly different result.

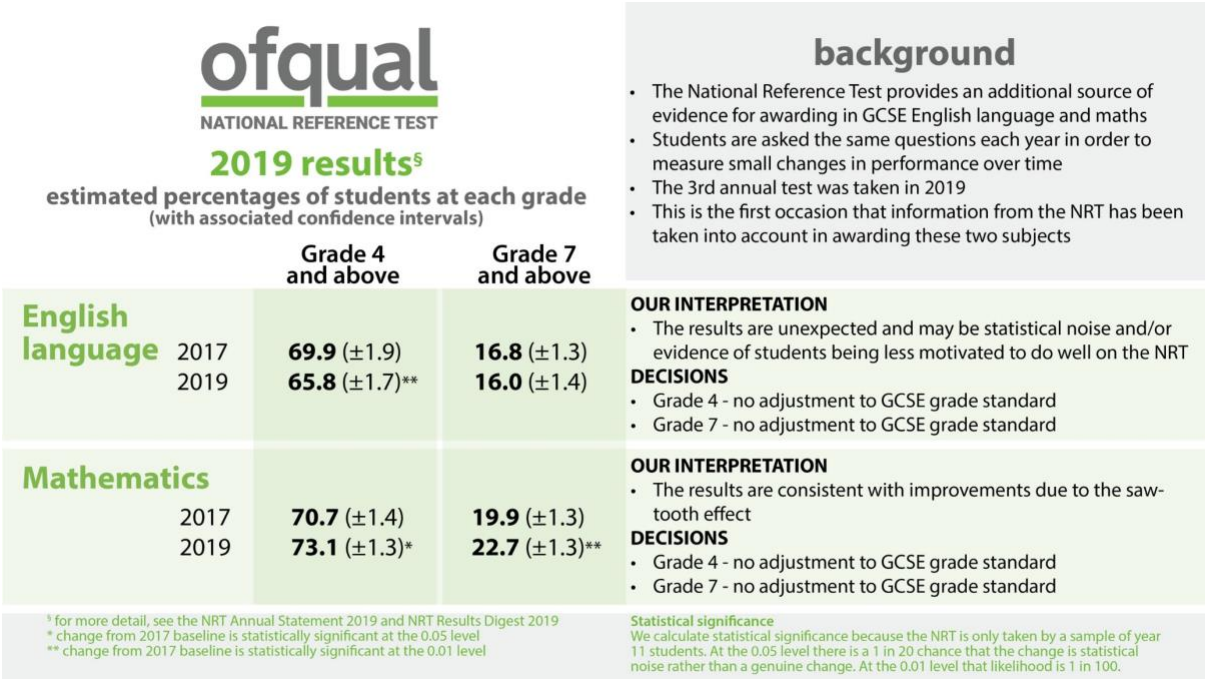
The diagram below shows the changes in the expected percentage of students at the grade 7, grade 5 and grade 4 boundaries, compared to 2017. The expected percentages are generally up in maths and slightly down in English for two of the three grade boundaries.

In English, NFER report a statistically significant downward change at grade 4 only. This is unexpected, and not consistent with the small improvement we might expect to see in the first years of a new qualification, as teachers become more familiar with the requirements (known as the 'sawtooth' effect).

In maths, NFER report a statistically significant upward change at grade 7, which suggests that student performance has improved slightly. This is not surprising, as we might expect to see an improvement in performance in the first years, as teachers get more familiar with the requirements of the new GCSEs.

The diagram below shows the results in 2017 and 2019, as well as the confidence intervals around those figures. Although the percentage of students at those grades in 2017 is fixed, the confidence intervals reflect the reported precision of the 2017 NRT results.

The diagram below also shows our decisions in relation to each of the grades in GCSE English language and maths.



ofqual

NATIONAL REFERENCE TEST

2019 results§

estimated percentages of students at each grade
(with associated confidence intervals)

		Grade 4 and above	Grade 7 and above
English language	2017	**69.9** (±1.9)	**16.8** (±1.3)
	2019	**65.8** (±1.7)**	**16.0** (±1.4)
Mathematics	2017	**70.7** (±1.4)	**19.9** (±1.3)
	2019	**73.1** (±1.3)*	**22.7** (±1.3)**

§ for more detail, see the NRT Annual Statement 2019 and NRT Results Digest 2019
* change from 2017 baseline is statistically significant at the 0.05 level
** change from 2017 baseline is statistically significant at the 0.01 level

background

- The National Reference Test provides an additional source of evidence for awarding in GCSE English language and maths
- Students are asked the same questions each year in order to measure small changes in performance over time
- The 3rd annual test was taken in 2019
- This is the first occasion that information from the NRT has been taken into account in awarding these two subjects

OUR INTERPRETATION
- The results are unexpected and may be statistical noise and/or evidence of students being less motivated to do well on the NRT

DECISIONS
- Grade 4 - no adjustment to GCSE grade standard
- Grade 7 - no adjustment to GCSE grade standard

OUR INTERPRETATION
- The results are consistent with improvements due to the saw-tooth effect

DECISIONS
- Grade 4 - no adjustment to GCSE grade standard
- Grade 7 - no adjustment to GCSE grade standard

Statistical significance
We calculate statistical significance because the NRT is only taken by a sample of year 11 students. At the 0.05 level there is a 1 in 20 chance that the change is statistical noise rather than a genuine change. At the 0.01 level that likelihood is 1 in 100.

Using the NRT evidence in awarding

GCSE awarding is guided by statistical predictions based on the prior attainment of the cohort, with input from senior examiners looking at the quality of students' work and comparing it to work in previous years. NRT results provide an additional source of evidence in the awards for GCSE English language and GCSE maths.

We have always been clear that we would not consider NRT evidence in GCSE awarding until 2019, because any improvement before then was likely to be due to the sawtooth effect, as teachers get used to the new GCSEs.

This summer, we discussed the results with exam boards in June, before deciding whether to make any adjustments to GCSEs in English language and maths.

In considering the evidence from the NRT, we aim to make sure that:

- our decisions are consistent over time and between subjects, regardless of the direction of any change
- we take account of contextual evidence from the student survey and other sources, and that we act cautiously in making any adjustments to grade standards
- we document and publish the reasons for our decisions.

In order to help us interpret the NRT results, we carry out additional analysis to consider the prior attainment profile of the sample of students who take the test. We also consider the findings from the student survey in relation to student motivation and students' views of the importance of the NRT and GCSE in English language or maths.

Prior attainment profile of the sample

In both English and maths, the achieved sample – that is, those students who took the test as opposed to all those who were selected to take part – has an upward bias in terms of prior attainment, demonstrated by the difference in the Key Stage 2 profile of the drawn and achieved samples. This is not, in itself, problematic. This difference was also seen in 2017 and 2018, and this bias has remained stable across the three years of the NRT. There is, therefore, no reason to believe that the bias in the achieved sample is having any material impact on the changes in results between 2017 and 2019.

Student motivation

Immediately after taking the NRT, students also take a short survey to capture, among other things, their NRT-specific test motivation, preparation for GCSEs, and motivation, feelings and attitudes about learning the relevant GCSE subject. The aim of the survey is to provide context for any changes in NRT results. The survey was introduced in 2017 and was also administered in 2018 and 2019.

For English, there were a number of small changes in the survey results for 2019, which together might explain some of the drop in results. Students reported slightly lower effort on the NRT, lower perceived importance of the NRT, greater indifference to their own NRT performance, less test preparation, less outside-school tuition, and lower views of the utility value and importance of the subject. Based on previous survey results, these small changes might be expected to produce slightly lower NRT results. These factors need to be taken into account when interpreting the NRT results as some might lead us to question whether the changes in NRT results would be reflected GCSE performance.

For maths, there have been fewer across-year changes in the survey results. Students reported lower perceived importance of the NRT, greater indifference to their own NRT performance, and increased enjoyment of the subject. None of these changes would lead us to expect any changes in maths results, on the basis of previous survey results.

It seems likely, therefore, that students taking the English NRT in 2019 were less motivated to do well in the NRT and also less enthusiastic about English more generally, but it is not possible to quantify the potential impact of this. Lower motivation might explain some of the apparent decline in results, but, given that the changes in survey results were small, it seems unlikely that it would explain the entirety of the change in English results.

Interpreting the results

In interpreting the results from this year's NRT, we considered carefully the threshold we should use for determining statistical significance. NFER report statistical significance at the 0.05 and 0.01 levels, and we have decided that we will focus on the 0.01 level, due to the high stakes nature of the test and GCSE results.

We have also considered carefully the statistical correction required to take account of the multiple comparisons we are making across years. As the number of comparisons increases, so does the likelihood that one will be statistically significant, purely by chance. As we agreed with NFER, they have compared each year with all other years (nine comparisons) and this is reported in the Results Digest. However, this year, we are specifically interested in whether the change from 2017 to 2019 is statistically significant. We therefore need to correct for a smaller number of comparisons (three comparisons). Hence fewer significant differences are reported in the NFER Results Digest than in this statement. Note that this is only concerned with the statistical significance of the change in outcomes between years, and does not have a bearing on the actual NRT outcomes themselves.

Our rationale is set out in more detail in Annex 3.

English

Having considered the evidence and the principles set out above, we believe there could be a case to make a small downward adjustment to the grade standards (which would tend to mean slightly higher grade boundaries) at grade 4 for 2019, particularly since the trend in the NRT results is counter to what we might expect to see as a result of the sawtooth effect in the first two or three years of a new qualification.

However, we have always been clear that we would be cautious in using this evidence. For us to make an adjustment in only the third year of the test, based on results in English which are unexpected, we would want to be confident that we were not at risk of interpreting statistical noise and/or NRT behavioural change as a real change in the anticipated GCSE performance. In future years, with more years of NRT data, we will be able to make more informed judgements about what the results show. Given that we could not be confident that these results would actually be reflected in poorer performance in GCSE English language, we decided not to make any adjustments in English for 2019.

Maths

Having considered the evidence and balanced the principles set out above, we did not believe there was a sufficiently strong case for making an adjustment at grade 7 in maths in 2019. We believe that the increase at grade 7 is consistent with the increase that we might expect to see as a result of the sawtooth effect in the first two or three years of a new qualification. Therefore we did not make any adjustment in maths for 2019.

We provided detailed NRT briefing documents to all senior examiners involved in awarding GCSE English language and GCSE maths in summer 2019 (see Annexes 1 and 2). However, this was one of a number of sources of evidence used in those awards. As in previous years, senior examiners will have been guided by predictions based on the prior attainment of each exam board's cohort, and they will have looked at student work from this year and last year.

Further information

There is more information about the NRT in the Background Report and NFER Results Digest.

Annex 1

ofqual

National reference test – briefing note for English language awarders

Each year students in a sample of approximately 300 schools in England sit the National Reference Test (NRT) in English and maths. The NRT is designed to measure changes in students' performance over time in GCSE English and maths, which might otherwise be difficult to detect. The NRT uses the same questions year-on-year and so we can be confident that any changes in performance are not due to a particular year's paper being more or less demanding. The NRT questions are based on the sorts of questions used in GCSE English language and maths and cover all of the content.

Results are reported at 3 key grade boundaries – 7/6, 5/4 and 4/3 – and show the estimated percentage of students expected to achieve those grades (and above) set against the baseline cohort in 2017. The results are reported with confidence intervals – these reflect the possibility that we might have got slightly different results if a different sample of students had taken the NRT. If the difference in results is larger than the confidence interval, we can generally be more confident that it reflects a real difference in performance between years. Statistical testing is undertaken to consider whether any differences in the percentage of students at each grade (and above) between years are significant.

If we believe that the NRT evidence is sufficiently compelling to make an adjustment at one or more grades, we will do this by making an adjustment to the predictions that guide awarding. We have agreed with exam boards that in principle we would not normally make an adjustment to grade 5, given that grade 5 is an arithmetic grade. This briefing therefore focuses on grades 7 and 4.

What do the NRT 2019 results show?

Table 1 and Figure 1 show the estimated percentage of students on the English language NRT at each grade (and above) and the confidence intervals around these estimates. Figure 1 illustrates that there has been a decline in the estimated percentage of students achieving at grades 7, 5 and 4 (and above) since the baseline in 2017. The key question in summer 2019 is whether any changes from previous years represent a genuine change in attainment.

We compared the estimated percentage of students at each grade in 2019 with the baseline year in 2017. This showed that the change was statistically significant at grade 4 at the 0.01 level of significance.[1] The change at grade 7 was not statistically significant.

[1] There are different levels of statistical significance. A 0.05 significance level indicates a 1 in 20 chance of the difference occurring by chance; at the 0.01 level of significance, that reduces to a 1 in 100 chance.

Table 1 Estimated percentages at each grade – English

Threshold	Grade 4 and above	Grade 5 and above	Grade 7 and above
2017	69.9 (68.0-71.8)	53.3 (51.4-55.2)	16.8 (15.5-18.1)
2018	68.8 (66.8-70.8)	52.8 (50.7-54.9)	16.8 (15.4-18.2)
2019	65.8 (64.1-67.5)	49.8 (47.8-51.8)	16.0 (14.6-17.4)

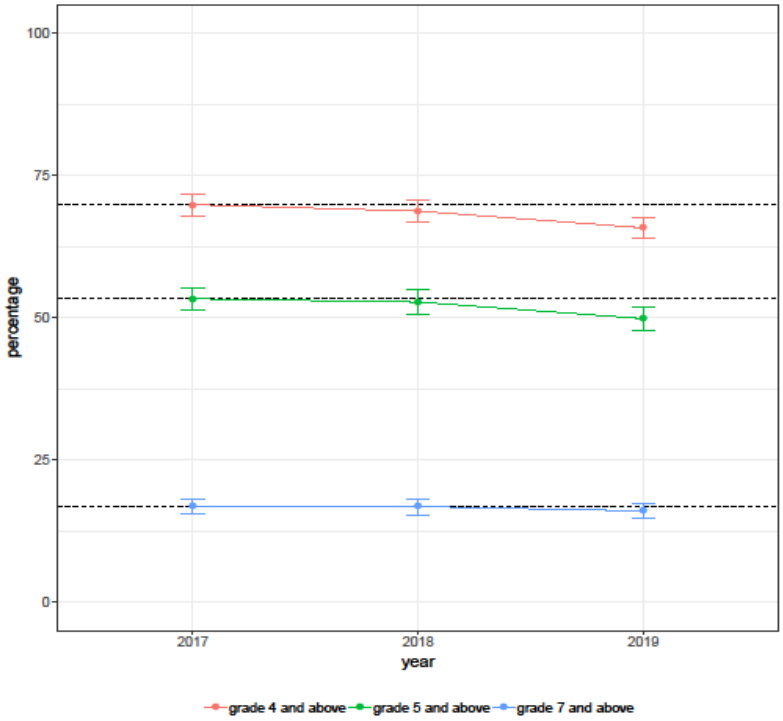


Figure 1 Estimated percentage of students at each grade with trendlines – English

What else do we know about the NRT 2019?

Each year we collect contextual information to help us interpret the results of the NRT. This includes information about the sample of students taking the test, their motivation, the quality of marking for the test and any other evidence that is available.

Our analysis shows that in 2019 the sample of test-takers (for English language) was similar to previous years. The student survey, however, indicated that there was lower reported test-taking effort, lower perceived importance of the NRT and greater indifference to students' own results of the NRT. While these changes point towards poorer NRT performance in 2019, the changes were relatively minor. There was also some evidence of a slight tendency towards more severe marking in 2019. Again,

however, the changes were small. We would not expect the changes in test motivation and marking to result in any large changes in NRT outcomes.

We also consider any other factors that might influence the results in a given year. This year is the third sitting of the NRT and we know that in the early years that a new assessment is available performance is likely to improve as teachers become more familiar with the new requirements. This is known as the 'sawtooth' effect.[2] The sawtooth effect means that in the early years that a new assessment is available, performance will generally look to have improved. However, this is because of increasing familiarity with the test, rather than a genuine improvement in attainment. We do not think it is fair to reward changes in performance resulting from the sawtooth effect since to do so would disadvantage students entering in the first year that a new assessment is available. This is why we prioritised statistical predictions in the first two years of awarding the reformed GCSE qualifications.

The NRT questions are similar to questions in the reformed GCSEs, so we might expect performance to improve in the NRT over the first few years due to a sawtooth effect. This is not what we saw in English language and the NRT English results are therefore unexpected in the context of a likely sawtooth effect. However, while our research suggests that the sawtooth effect is a generally observed phenomenon, we do not know whether it affects some subjects more than others or indeed the exact size of the effect in individual subjects.

What is Ofqual's decision?

It is for Ofqual to consider the results of the NRT and decide whether an adjustment should be made, having allowed feedback from exam boards on our proposed approach. When making our decision, we take account of the results of the test and the contextual information.

Taking account of the available information, Ofqual have decided that in 2019 there will be no adjustment at either grade boundary (7/6 or 4/3) for English language. If we were to make an adjustment, we must be confident that the change in NRT results represents a genuine change in the attainment of students, rather than statistical noise, or other factors. We are not confident that that is the case this year.

What does this mean for awarders' roles at the awarding meeting?

The role of awarders is to scrutinise student work to ensure that the performance in a given year is comparable to previous years – thereby maintaining performance standards from one year to the next. We know, however, that it is very difficult for awarders to make precise judgements at awarding (eg between adjacent marks) and there is a body of research evidence illustrating this. The purpose of the NRT is to detect changes in performance that would be difficult to detect judgementally. Where we think that attainment has genuinely improved, we will make an adjustment to the statistical predictions that guide the awards.

The purpose of making an adjustment is to ensure that students' are appropriately rewarded for their performance. This means that, if an adjustment is made, the work that awarders are asked to scrutinise should be of the same performance standard as the archive work from previous years. The task being asked of awarders is therefore no different from other years in a stable qualification. Equally, if an

[2] https://www.gov.uk/government/publications/investigation-into-the-sawtooth-effect-in-gcses-as-and-a-levels

adjustment is not made, this is because we do not believe there has been a genuine change in attainment. Therefore, the work that awarders are asked to scrutinise should represent a comparable performance standard to previous years.

While we expect performance to look similar to previous years in either case, we are also relying on awarders to identify instances where this is not so (within the limits of making such judgements). In such instances, we would require awarders to capture the appropriate evidence for Ofqual to consider.

Annex 2

ofqual

National reference test – briefing note for maths awarders

Each year students in a sample of approximately 300 schools in England sit the National Reference Test (NRT) in English and maths. The NRT is designed to measure changes in students' performance over time in GCSE English and maths, which might otherwise be difficult to detect. The NRT uses the same questions year-on-year and so we can be confident that any changes in performance are not due to a particular year's paper being more or less demanding. The NRT questions are based on the sorts of questions used in GCSE English language and maths and cover all of the content.

Results are reported at 3 key grade boundaries – 7/6, 5/4 and 4/3 – and show the estimated percentage of students expected to achieve those grades (and above) set against the baseline cohort in 2017. The results are reported with confidence intervals – these reflect the possibility that we might have got slightly different results if a different sample of students had taken the NRT. If the difference in results is larger than the confidence interval, we can generally be more confident that it reflects a real difference in performance between years. Statistical testing is undertaken to consider whether any differences in the percentage of students at each grade (and above) between years are significant.

If we believe that the NRT evidence is sufficiently compelling to make an adjustment at one or more grades, we will do this by making an adjustment to the predictions that guide awarding. We have agreed with exam boards that in principle we would not normally make an adjustment to grade 5, given that grade 5 is an arithmetic grade. This briefing therefore focuses on grades 7 and 4.

What do the NRT 2019 results show?

Table 1 and Figure 1 show the estimated percentage of students on the maths NRT at each grade (and above) and the confidence intervals around these estimates. Figure 1 illustrates that there has been a rise in the estimated percentage of students achieving at grades 7, 5 and 4 (and above) since the baseline in 2017. The key question in summer 2019 is whether any changes from previous years represent a genuine change in attainment.

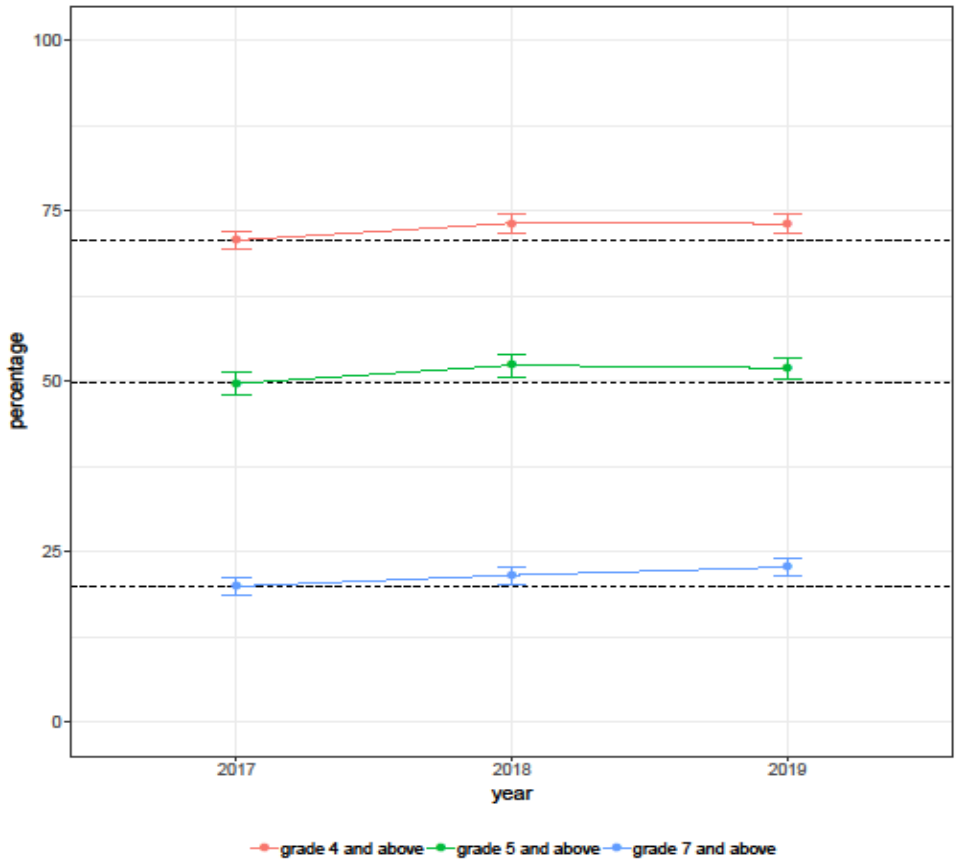
We compared the estimated percentage of students at each grade in 2019 with the baseline year in 2017. This showed that the changes were statistically significant at grade 7 at the 0.01 level of significance.[3] The changes at grade 4 were statistically significant at the 0.05 significance level.

[3] There are different levels of statistical significance. A 0.05 significance level indicates a 1 in 20 chance of the difference occurring by chance; at the 0.01 level of significance, that reduces to a 1 in 100 chance.

Table 1 Estimated percentages at each grade – maths

Threshold	Grade 4 and above	Grade 5 and above	Grade 7 and above
2017	70.7 (69.3-72.1)	49.7 (48.0-51.3)	19.9 (18.6-21.2)
2018	73.2 (71.7-74.7)	52.4 (50.7-54.0)	21.5 (20.2-22.8)
2019	73.1 (71.8-74.4)	51.9 (50.3-53.5)	22.7 (21.4-24.0)

Figure 1 Estimated percentage of students at each grade with trendlines – maths



What else do we know about the NRT 2019?

Each year we collect contextual information to help us interpret the results of the NRT. This includes information about the sample of students taking the test, their motivation, the quality of marking for the test and any other evidence that is available.

Our analysis shows that in 2019 the sample of test-takers and their NRT test taking effort (in maths) were similar to previous years. There was some decline in the perceived importance of the NRT and increased indifference to students' own NRT results. Our analysis also shows that the marking was consistent with previous

years. Overall, this suggests that the context within which the test has operated has remained broadly stable.

We also consider any other factors that might influence the results in a given year. This year is the third sitting of the NRT and we know that in the early years that a new assessment is available performance is likely to improve as teachers become more familiar with the new requirements. This is known as the 'sawtooth' effect.[4] The sawtooth effect means that in the early years that a new assessment is available, performance will generally look to have improved. However, this is because of increasing familiarity with the test, rather than a genuine improvement in attainment. We do not think it is fair to reward changes in performance resulting from the sawtooth effect since to do so would disadvantage students entering in the first year that a new assessment is available. This is why we prioritised statistical predictions in the first two years of awarding the reformed GCSE qualifications.

The NRT questions are similar to questions in the reformed GCSEs, so we might expect performance to improve in the NRT over the first few years due to a sawtooth effect. We would not make an adjustment if we thought an improvement was due to a sawtooth effect, since this reflects increased familiarity with the assessments, rather than a genuine improvement in attainment.

What is Ofqual's decision?

It is for Ofqual to consider the results of the NRT and decide whether an adjustment should be made, having considered feedback from exam boards on our proposed approach. When making our decision, we take account of the results of the test and the contextual information.

Taking account of the available information, Ofqual have decided that in 2019 there will be no adjustment at either grade boundary (7/6 or 4/3). This is because we think that the evidence points to the improvement in maths being a result of the sawtooth effect rather than a genuine improvement in attainment.

What does this mean for awarders' roles at the awarding meeting?

The role of awarders is to scrutinise student work to ensure that the performance in a given year is comparable to previous years – thereby maintaining performance standards from one year to the next. We know, however, that it is very difficult for awarders to make precise judgements at awarding (eg between adjacent marks) and there is a body of research evidence illustrating this. The purpose of the NRT is to detect changes in performance that would be difficult to detect judgementally. Where we think that attainment has genuinely improved, we will make an adjustment to the statistical predictions that guide the awards.

The purpose of making an adjustment is to ensure that students' are appropriately rewarded for their performance. This means that, if an adjustment is made, the work that awarders are asked to scrutinise should be of the same performance standard as the archive work from previous years. The task being asked of awarders is therefore no different from other years in a stable qualification. Equally, if an adjustment is not made, this is because we do not believe there has been a genuine

[4] https://www.gov.uk/government/publications/investigation-into-the-sawtooth-effect-in-gcses-as-and-a-levels

change in attainment. Therefore, the work that awarders are asked to scrutinise should represent a comparable performance standard to previous years.

While we expect performance to look similar to previous years in either case, we are also relying on awarders to identify instances where this is not so (within the limits of making such judgements). In such instances, we would require awarders to capture the appropriate evidence for Ofqual to consider.

Annex 3: Interpreting statistical significance

Determining whether the changes in NRT outcomes are statistically significant is not straightforward. Quantitative research[5] distinguishes between type I errors (in this case, wrongly believing that statistical noise is a change in student performance) and type II errors (in this case, wrongly dismissing a real change in performance as statistical noise). Determining statistical significance will partly depend on the relative importance of avoiding a type I or a type II error[6].

In the context of the NRT, there are risks around making an adjustment where we are not completely confident that the change represents a real change in performance, as this would undermine public confidence and our statutory objective to maintain standards and would be unfair to students. In this first year of using the NRT evidence in awards, this risk is particularly acute. Given the high stakes nature of GCSE grades for students and schools, we believe it is important to be cautious in interpreting changes in NRT outcomes as representing a real change in performance. These considerations were behind the 'judicious' principle (outlined in paragraph 6c) as agreed by the Ofqual board in advance of the current data being considered.

Given that in summer 2019 there are multiple years' data, it is possible to make multiple comparisons (eg comparing 2017 and 2018, 2018 and 2019, or 2017 and 2019) across the three grades (7, 5 and 4). Where multiple comparisons are made, this must also be taken into account when judging statistical significance. As such, a correction has to be made to take account of the fact that as the number of comparison increases, so does the likelihood of finding a statistically significant difference purely by chance (making a type I error). There are different ways to do this. For the NRT, a relatively conservative approach has been used – Bonferroni, which is also used by large scale international surveys such as PISA. However, a judgement has to be made about which comparisons are the most pertinent in any given year. As agreed, NFER have made a correction for nine comparisons (three year-on-year comparisons at each of the three grades). In future years, if we continue to compare each year with every other year, the number of comparisons will increase (18 in 2020, 30 in 2021). The critical value for determining statistical significance will also increase, which will increase the risk of making a type II error (dismissing a real change as statistical noise).

We have reflected on this risk and concluded that it would be more appropriate to compare the current year with the 2017 baseline, for each of the three grade boundaries.

Taking a highly conservative approach to correcting for multiple comparisons would make it very unlikely that a real change would be detected (increasing the risk of a

[5] See for example: Clark-Carter, David. *Quantitative psychological research: a student's handbook* 3rd ed (2010)
[6] Through the setting of alpha e.g. at 0.05, 0.01 or 0.001 levels

type II error). A more moderate approach which recognises the context in which the comparisons are made would seem more appropriate. As a result, fewer significant differences are reported in the NFER Results Digest than in this statement.

This approach fits with what we believe to be the most meaningful comparison this year: to compare 2019 (and in future, the year under consideration) with the baseline established in 2017. This appropriately balances the risk of a type I versus type II error given the high stakes nature of GCSE grades for students and schools.

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

22 August 2019

Ofqual/19/6527/1