

RESEARCH AND ANALYSIS

Maintaining standards in summer 2019

ofqual

Contents

Introduction	3
Our approach to maintaining standards	3
<i>Research to improve the quality of examiner judgments in awarding</i>	5
<i>National Reference Test</i>	6
<i>Inter-board comparability</i>	9
Specific issues in 2019	12
<i>Reformed A level maths</i>	12
<i>Statistical predictions for AS in 2019</i>	16
<i>First awards of Pearson Level 1/2 Tech awards</i>	17
Conclusion and next steps	18
Appendix A: The use of statistics in awarding	19

Introduction

We monitor GCSE, AS and A level awards each summer to make sure, as far as is possible, that there is a level playing field for students and that they get the grades that their performance deserves. We do this so that standards are maintained over time and between exam boards, and in particular so that it is not easier or more difficult to get a particular grade in a subject with one exam board than with another.

In summer 2019 we monitored the awards in over 250 GCSEs, AS and A levels. We also monitored a number of vocational and technical qualifications that are included on Department for Education (DfE) school and college performance tables as Technical Awards, Technical Certificates, Applied Generals and Tech Levels.

This report explains the work we do to make sure that standards are set and maintained appropriately, our ongoing work to incorporate more reliable examiner judgements into the decisions about where grade boundaries are set, and analysis to provide context to the National Reference Test (NRT) results.

We also report on a number of particular cases from the summer. This includes the first full awards of reformed A level maths, an error that came to light in September in relation to the predictions used for AS, and a change to the grade point thresholds in the first awards of new BTEC Level 1/2 Tech Awards.

Our approach to maintaining standards

We expect exam boards to maintain standards from one year to the next, so that what students must demonstrate to achieve a particular grade does not vary significantly over time. There will of course be small variations, depending on the particular questions asked and the overall demand of the paper, but those should not make it substantially easier or harder to achieve a grade one year compared to another.

The demand of papers from one year to the next does inevitably vary, and grade boundaries are set to take account of that. If one year's paper is more demanding, this will make it harder for students to achieve marks than in previous years, and so grade boundaries will tend to be lower. Conversely, if one year's paper is easier than in previous years, the grade boundaries will tend to be higher.

Making precise judgements about where to set grade boundaries is challenging, even for experienced senior examiners. While they might be able to make broad judgements about the standard of work that, for example, a GCSE grade 4 or grade 7 represents, making precise judgements about which of 2 adjacent marks better represents the standard at that grade is more challenging. One of the reasons for that is, in general, similar student performance on an easier paper will tend to

look better than student performance on a harder paper. Another reason is that students can achieve their marks in many different ways, so two students with the same total number of marks on a paper might each have performed very differently across that paper.

Exam boards use predictions based on the prior attainment of the cohort (Key Stage 2 for GCSE, GCSE for AS and A level) to guide these fine judgements. In general, predictions give us a good estimate of where grade boundaries should be set to maintain standards from one year to the next, so that what students have to do to achieve a particular grade one year is broadly the same as students in previous years.

Predictions are also useful where the cohort taking a subject has changed, perhaps because many more students are taking a subject, or because entry patterns have changed. Crucially, predictions are the only tool we have to make sure that standards are aligned between exam boards in a subject, so that it is no easier to achieve a grade with one board than with another. This is vital, in order that the system is fair for all students.

Over the last 10 years or so, the process of setting grade boundaries has changed very little. Where there have been changes, they have generally been to move face-to-face meetings online, so that awarders review student work onscreen and then come together in a virtual environment to discuss and agree final grade boundary recommendations.

Currently, awarding meetings (whether face-to-face or virtual) involve a balance of examiner judgment about the quality of student work compared to previous years, and statistical predictions about students' expected results, given their prior attainment profile. Examiners' judgements about where to set grade boundaries can, therefore, be influenced by the statistical evidence.

In the move to reformed GCSEs and A levels, our approach has been to prioritise statistics in order to minimise any disadvantage to those students who were the first to sit the new qualifications. Our previous research into the sawtooth effect¹ suggested that students might perform less well in the first few years of a new qualification. This, and the differences in subject content, structure and assessment between old and new specifications, can make it much more difficult to make precise judgements about the quality of student work.

Now that the transition to reformed qualifications is largely complete, we are keen to make sure exam boards are able to detect any changes in student performance over time. In a 'steady state' exam boards can have more confidence in the judgements of

1

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549686/an-investigation-into-the-sawtooth-effect-in-gcse-as-and-a-level-assessments.pdf

their senior examiners, but we need to balance that with making sure that standards in a subject are aligned between exam boards.

In GCSE English language and maths we have evidence about changes in student performance over time from the National Reference Test (NRT). We have also been working with exam boards to research different approaches, for all subjects, to collect examiners' judgement in ways that are independent of the statistical evidence and less prone to bias.

Research to improve the quality of examiner judgments in awarding

This work has involved comparative judgement and/or rank ordering techniques. We have published a [separate report with more detail about the experimental work carried out in 2018 and 2019](#). This work was carried out alongside or shortly after live awarding and was intended to explore the feasibility and manageability of the approach, as well as the plausibility of the grade boundary estimates that were generated from the judgements.

Comparative judgement/rank ordering

Comparative judgement (CJ) is a well-researched technique in assessment. First identified by Thurstone (1927)² it involves comparing 2 or more objects on the basis of a single attribute (for example, quality). Repeated comparisons of different pairs among a set of objects allows the construction of a single trait scale and the location of each object on that scale ('measure of perceived quality').

CJ is potentially a natural fit to the purpose of awarding/setting grade boundaries in that it allows direct comparison of the quality of student work from 2 sessions – the previous year and the current (live) session.

CJ methods also have a number of advantages over the current approach where senior examiners aim to judge whether particular pieces of work are worthy of one grade or another. In CJ methods, senior examiners do not need to all have the same 'internal' standard because they are not making absolute judgements – they are simply comparing 2 pieces of work and deciding which is better. Whether a senior examiner tends to be severe or lenient has very little effect (unlike in the current system of making absolute judgments). Judgements are also made independently from statistics such as predictions (and knowledge of the marks on the scripts) and so this provides an independent source of evidence of the relative performance (across the mark range) in at least 2 years/sessions.

² Thurstone, L.L.. (1927), A law of comparative judgement. *Psychological Review* 3:273-86

The statistical modelling used to analyse results can help us evaluate the quality of the judgements, for example by providing data on the extent to which some judges have made different patterns of judgements from other judges.

There are various forms of CJ that can be used, including the following.

- paired comparative judgment (CJ) – repeated pairs of scripts (sometimes from the same year and sometimes from different years) are judged, with the ‘better’ script identified (for example ‘which one is the better quality script?’)
- rank ordering (RO) – groups of scripts from both years are placed in a single rank order according to quality
- “crowd sourcing” – instead of using a small panel of expert judges, this involves conducting a paired CJ with many judges (for example, teachers) making fewer judgements. The potential advantage of this is gaining a wider pool of expertise

In summer 2018 we piloted several different versions of CJ across 5 different specifications for 4 exam boards. We reviewed the evidence from those pilots and in summer 2019 we ran a further pilot in GCSE English language.

Overall, the results suggest that CJ methods are very promising for capturing expert judgement for the purpose of setting grade boundaries to maintain standards from one year to the next. The totality of the pilots indicate that pooling a sufficiently large number of judgements over most of the mark range can give reliable outcomes and potentially increase the validity of expert judgement in standard maintaining.

Exam boards have also been carrying out their own pilots and early in 2020 we will discuss with them the outcomes from those. There are a number of considerations that we will need to bear in mind, including the merits of different designs in operational contexts, and the relative weight such methods might carry in relation to other sources of evidence, in particular the statistical predictions already used.

National Reference Test

The National Reference Test (NRT) is taken every February/March by a nationally representative sample of year 11 students. It provides an additional source of information that can be used in GCSE English language and maths awarding later in the summer. The NRT is administered for Ofqual by the National Foundation for

Educational Research (NFER), and they report annually on the operation of the test and the test results.³

In summer 2019 we decided to make no adjustment to GCSE grades as a result of NRT evidence. We judged that the small increase in performance in maths was consistent with the pattern we might expect in the first years of a new qualification as schools become more familiar with the requirements. In English, we were not sufficiently confident that the decrease in NRT performance would be reflected in lower performance in GCSE.

In coming to that decision, we reviewed data from a number of sources, in addition to the NRT results reported by NFER. The additional analyses help us to contextualise and interpret the annual NRT results. Ahead of the decision about whether or not to adjust GCSE grade standards, we consider the results of the NRT student survey into students' motivation and attitudes towards NRT and GCSE over time, and we carry out analysis of the students taking the NRT compared to those in previous years. After results are issued, we also carry out further analysis to look at the relationship between NRT performance and GCSE results, for those students who took the NRT. This enables us to evaluate the effectiveness of the NRT as a predictor of GCSE results.

We have published a [separate report detailing the results of these analyses for the NRT in 2017 to 2019](#). The following sections provide a brief summary of the findings.

Comparison of students taking NRT with all year 11 students

In each school that is selected to take the NRT, up to 30 students are randomly selected to take the English test and another 30 to take the maths test. Inevitably, there will be some absences, as on any given day there will be students who are unwell or do not attend school. The actual number of students who take the NRT on the agreed date in a given school is therefore often lower than 60. We are interested to know whether there are any differences between students who sit the NRT and those who are selected to sit it but do not, for whatever reason, take part.

Our analysis shows that in all 3 years of the NRT (2017, 2018 and 2019) students with lower prior attainment at Key Stage 2, and those subsequently achieving lower GCSE grades were more likely to be absent from the NRT. This suggests a slight upward bias in NRT results because the sample includes slightly fewer lower ability students. However, this effect has been stable across the 3 years of the NRT, which means that the slight upward bias in NRT results has not fluctuated from year to

³ Ofqual has published NFER's annual NRT Results Digest since 2018. The documents can be accessed at <https://www.gov.uk/government/collections/national-reference-test-information>

year. Therefore, in 2019 we could be confident that any changes in NRT results were not due to changes in the students sitting the test.

Student survey

Following the NRT test, students are given a short survey to complete.

The first section of the student survey asks about motivation on the NRT. In both English and maths, students taking the NRT in 2019 perceived it as less important and reported less NRT-specific preparation and more indifference to their own NRT performance than their 2017 counterparts. The 2019 students also reported less effort made in the NRT than the 2017 participants, and in English this decrease from 2017 was statistically significant (that is, it was unlikely to have occurred by chance).

The second section of the survey asks students about their preparation for GCSE. In English, the questions are about the type of English lessons received at school (combined, separate English language and literature lessons, or language lessons only), teaching time, and any tuition received outside school. Overall the responses were similar in 2019 to previous year, with the only statistically significant change being a smaller proportion of participants who reported receiving tuition outside school in 2019 than in 2017.

In maths, the questions are about teaching time, homework time, time for maths-related activity at school outside formal lessons, and any tuition received outside school. Overall, the responses were similar to those in 2017.

The final section of the survey asks students about their motivation in relation to the relevant subject at GCSE. For English, the 2019 NRT students reported that they found the subject less useful and less important, and saw less of a role for the subject in their future plan than their 2017 counterparts. For maths, the 2019 NRT students reported more enjoyment of the subject than their 2017 counterparts.

In June, we concluded that the small decrease in motivation reported by students taking the English NRT was unlikely to account for all of the change in NRT English results, but we were not confident that the decrease in NRT performance would also be seen in GCSE in August. We therefore did not believe that an adjustment was justified. Our rationale for this decision is set out in more detail in our Annual Statement⁴ published in August.

4

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826570/NRT_annual_statement_2019_-_FINAL196527.pdf

Comparison of NRT performance with GCSE grades

For the NRT to provide useful information for GCSE awarding, students' performance in the NRT should, ideally, predict their own and their cohort's GCSE performance. However, a number of factors mean we do not necessarily expect a perfect relationship between NRT and GCSE performance.

First, the NRT is dissimilar in length, and also, to a lesser extent, in the exact question style and format, to the GCSE exams that students will have been preparing for, and these differences might affect students in different ways. In addition, GCSE exams take place some months after the NRT, and students taking the NRT might improve their knowledge and skills to different degrees in the intervening months. A further difference is that the NRT is relatively low-stakes for individual students and schools (whereas GCSEs are high-stakes) and this can affect student motivation and effort. Any change in the relationship between NRT and GCSE performance over the years would signal a possible change in the effect of some or all of these differences on NRT test performance.

In the autumn following GCSE results, for each of the 3 key grade boundaries – 7/6, 5/4 and 4/3 – we look at whether students perform similarly on NRT and GCSE and if not, whether they perform better on NRT or GCSE.

In general, the proportion of students who perform similarly on both has been highly stable across the 3 years of the NRT. Where student performance differs, the pattern has changed over time. In both English and maths there are more students who perform better on GCSE than on NRT. Comparing 2019 with 2017, this proportion has increased in English but decreased in maths. This suggests that, to a small extent, the NRT results under-predict GCSE performance in English and that this is increasing over time.

We are currently planning more work to try to explain the unexpected results in the NRT in English. This will include an analysis of any changes over time in student performance on particular questions or areas of the subject content in both the NRT and GCSE. We will report on that work in 2020.

Inter-board comparability

This is an important aspect of ensuring fairness for students, as well as for schools and colleges. In any one year, students will be competing for sixth form or higher education places with others who may have taken the same subjects with different exam boards. If it is easier or harder to obtain a particular grade with one board compared to another, that is simply not fair to students. So our priority when monitoring the way exam boards set grade boundaries is to make sure it is no easier to obtain a particular grade with one exam board than another. Predictions enable us

to judge that, even when there are large changes in entry patterns, because they take account of any changes in the prior attainment profile of the entry.

GCSE, AS and A level

In GCSEs, AS and A levels, we put rules in place specifying that all qualifications with the same title must include the DfE content⁵. Our rules also specify particular aspects of the assessment (whether the assessment should be tiered, the weighting of any non-exam assessment, etc). The aim is that each exam board's specification will have similarly demanding content and assessment. Our expectation, therefore, is that it should be no easier to get a particular grade with one exam board than another.

Where entries are sufficiently large, we use statistical predictions to judge the comparability of grade standards across all exam boards in a subject. Where all exam boards' awards in a subject are reasonably close⁶ to their predictions, we judge that their grade standards are aligned, and therefore it is not easier or more difficult to get a particular grade with one board than another. Where the entry numbers are relatively small, exam boards will balance the use of statistics with the judgements of their senior examiners.

Our view is that in summer 2019 standards were aligned between boards in a subject. We reported in August on the small number of awards that were out of tolerance.⁷

Applied Generals

Several awarding organisations introduced new or revised qualifications for first teaching in September 2016. They are more commonly known by the brand names (such as BTECs, Cambridge Technicals) or other titles⁸. The term 'Applied General' or 'Tech Level' refers to their categorisation in DfE 16-18 performance tables, which include qualifications which meet certain design rules. One of these rules is a requirement for at least 40% (Applied Generals) or 30% (Tech Levels) of the qualification to be external assessment. This doesn't necessarily mean exams – it can be practical work, or controlled assessment, provided it is set and marked by the awarding body. Another current requirement is the restriction that a student can only re-sit an externally-assessed unit twice.

⁵ For example: [GCSEs \(9 to 1\): requirements and guidance](#)

⁶ In general, we define this as being within the published tolerance of plus or minus 1, 2 or 3 percentage points (depending on entry size) of the prediction. However, there may be instances where we judge that an exam board is out of line with others, even though its award is within tolerance.

⁷ <https://www.gov.uk/government/publications/monitoring-gcse-as-and-a-level-exams-summer-2019>

⁸ See full list here <https://www.gov.uk/government/publications/16-to-19-qualifications-discount-codes-and-point-scores>

These qualifications are different from GCSEs and A levels in that there is no common specified content. However, there is potential for considerable overlap between the content covered by some qualifications in the same subject area. We therefore carried out work to look at the content coverage and comparability of standards in the external units in the 4 largest-entry subjects with qualifications offered by at least two awarding organisations.

The first external assessments were taken by students in January and June 2017. Four subjects with relatively large entries were applied science, business, health and social care, and ICT. These were offered by three awarding organisations – AQA, OCR and Pearson (although not all subjects were offered by all awarding organisations).

Research

We carried out 3 strands of work. First, a subject expert review of the subject content in the external units. Second, we collected subject expert judgements about the quality of work at the pass and distinction grades. Third, we conducted a statistical analysis of the probability of a student achieving a particular grade in an external unit, given their prior attainment and other factors potentially affecting achievement in level 3 qualifications. All of this work was carried out with a view to reporting on and, if necessary improving, the alignment across awarding organisations within a subject. We will report on this work in more detail in 2020 but we present a summary here.

Our review of the subject content suggested that in the 4 subjects – applied science, business, health and social care, ICT– there was sufficient overlap in the knowledge, skills and understanding in the externally assessed units within a subject for us to expect a degree of comparability between the grades awarded by different awarding organisations. This is especially important as the grades are used interchangeably by UCAS and higher education for selection purposes, as well as for school and college accountability purposes.

We then asked subject experts to compare the quality of student work in 2017 at pass and distinction grades in each of the 4 subjects. They used a CJ approach to compare a number of different pieces of work at pass and distinction, from different awarding organisations. They found some evidence of misalignment in all subjects, although to varying degrees.

Alongside this qualitative work, we carried out a statistical analysis of the probability of students with similar average GCSE grade obtaining a pass grade on the externally assessed units. This varied depending on the awarding organisation. This analysis confirmed the misalignment in standards in some of these external units, both in 2017 and 2018.

We also carried out an analysis of the factors that predict achievement in these external units. We found that prior attainment (average GCSE grade) was the best predictor of student achievement on the externally assessed units in Applied Generals. We also looked at how well average GCSE grade predicted achievement in Applied Generals compared to how well it predicts achievement at A level. We found that average GCSE grade is almost as good a predictor of achievement in the external units in Applied Generals as it is a predictor for A levels.

Monitoring

Our monitoring work is focused on creating a level playing field for students taking similar qualifications. Having found some evidence of misalignment of grade standards, we took action to start to address this in the summer 2019 awards. We created a national prediction matrix, based on the relationship between students' average GCSE grade and their achievement in the externally assessed units in 2018, and the awarding organisations used those matrices to predict likely outcomes in the externally assessed units in summer 2019. Where there was evidence of misalignment, we agreed that they would move closer to the national standard, balancing the need to align standards with the need to maintain standards over time.

We monitored the summer 2019 awards and collected data on how far each awarding organisation had moved towards the national standard. We also collected additional data following the issue of results, so that we could evaluate the impact of the national predictions. The analysis showed that, in 3 of the 4 subjects, there was better alignment between awarding organisations in summer 2019. In the fourth subject, there had been reasonable alignment already, and that continued.

Next steps

We will report the detail of research in 2020. In the meantime, we believe predictions are a useful tool to improve alignment between awarding organisations within a subject in Applied Generals. We are therefore discussing with awarding organisations how best to use predictions more broadly, for example in January unit awards, and in other subjects.

Specific issues in 2019

Reformed A level maths

In August⁹ we announced our intention to investigate the reasons for the systematic and relatively large differences in the grade boundaries set in 2018 and 2019 for the

⁹ <https://www.gov.uk/government/news/guide-to-as-and-a-level-results-for-england-2019>

reformed A level maths specifications. Boundaries in 2019 were, in general, substantially lower than in 2018.

We were clear at the time that we were confident in the 2019 awards, because the size of the entry (around 85,000 students) meant that the statistical predictions were very reliable and there was plenty of student work for senior examiners to scrutinise. In contrast, in the 2018 awards, there were only just over 2,000 students across 3 exam boards – mostly year 12 students – taking the reformed specification. This meant that the statistical evidence was less reliable. In addition, because these students tended to be very able mathematicians who were also studying further maths, they were clustered at the top grades, meaning that there was very little student work for senior examiners to review at the grade E boundary.

We wanted to understand why the grade boundaries were so different between the 2 years, once the full range of evidence from 2018 and 2019 was available. We have published a [report which sets out the work in detail](#).

We carried out 3 inter-related strands of work, as follows.

- strand 1 – a statistical analysis of the results in 2018 and 2019, considering the results for year 12 and year 13 students separately, and comparing them to the legacy qualifications
- strand 2 – a comparison of the difficulty of the 2018 and 2019 question papers, looking at judgements of question difficulty as well as data on how students performed on each of the questions
- strand 3 – a comparison of student performance across 2018 and 2019

Strand 1 – statistical analysis

In reformed qualifications, including for A level maths, our approach is to use statistics to carry forward the standard from the legacy qualifications, so that students are not disadvantaged by being the first to sit the new qualifications. In practice this meant using predictions to carry forward the ‘value-added’ relationship between GCSE and A level maths, so that, in general, a student who would have achieved a particular grade in the legacy qualification would achieve the same grade in the reformed qualification.

A level maths was unique in that we allowed a first award at the end of only one year of teaching the new specifications to allow the first cohort of students the opportunity to take A level maths at the end of year 12. We expected that the year 12 students entering at the end of that first year would likely be very able mathematicians who were also studying further maths. This proved to be the case.

Since the majority of students were in year 12, predictions in 2018 were based on the value-added relationship between GCSE and A level maths for year 12 students in the legacy qualification. In 2019, predictions were (as is usual) based on year 13 students, as they were by far the majority entry.

In the legacy qualifications, the profile of grades obtained by year 12 students was skewed towards the top grades, because they tended to be the very able students. The profile of grades for year 13 students was far less skewed. In general, the small minority of year 12 students entered for A level maths tend, on average, to achieve higher grades than the substantial majority of year 13 students for maths. This has always been the case.

Our analysis suggests that the gap between the achievement of year 12 and year 13 students has got wider in the reformed specifications, largely as a result of the move to linear qualifications and hence the removal of the opportunity to re-sit AS units in year 13. The move to linear had much less of an impact on year 12 students because in both legacy and reformed qualifications, they took all their exams at the end of year 12.

In the transition to the reformed qualifications, we said we would use statistical predictions to compensate for structural changes to qualifications, so that students were not disadvantaged by being the first to sit new qualifications. We estimate that this compensation in 2019 for the move to linear and the removal of re-sitting opportunities explains about half of the difference in the 2018 and 2019 grade boundaries.

Strand 2 – comparing the difficulty of the 2018 and 2019 papers

This strand of work looked at whether the lower grade boundaries were a result of more demanding papers in 2019. When grade boundaries are set, exam boards aim to take account of the relative difficulty of the papers, compared to previous years. If one year's papers are more difficult, then grade boundaries will be lower.

We carried out a CJ exercise in which we asked A level maths teachers to judge the difficulty of individual questions. This was similar to the work we did to compare the difficulty of questions in the specimen papers.¹⁰

We also looked at data about how students actually performed on the individual questions in 2018 and 2019. However, it is more difficult to draw conclusions about question and paper difficulty because of the different cohorts in 2018 and 2019. For example, 2018 questions might have appeared to be easier because they were taken by a more able cohort.

¹⁰ <https://www.gov.uk/government/publications/an-evaluation-of-the-item-difficulty-in-as-and-a-level-maths>

We combined the data from the CJ exercise with the question performance data to estimate the differences in difficulty between the 2018 and 2019 papers. In 3 out of the 4 specifications, this suggests that the 2019 papers were more difficult and therefore that grade boundaries ought to have been lower to take account of that.

Strand 3 – comparing student performance on reformed papers in 2018 and 2019

In this work, we looked at one paper per specification, choosing the paper which was judged to show the least difference in terms of expected difficulty between 2018 and 2019. Maths subject experts compared student performance, across the full grade range, on the 2018 and 2019 papers.

Making holistic judgements about the quality of the work across a whole paper can be challenging, particularly when judging work at grade E which tends to be more inconsistent. This work took into account these levels of uncertainty when drawing conclusions. A discernible difference in student performance between 2018 and 2019 was identified in 3 of the 4 specifications at grade A, with the same being true at grade E for 3 of the 4 specifications.

Summary and conclusions

This work has revealed a number of interacting factors which led to the relatively large changes in grade boundaries in the reformed A level maths specifications between 2018 and 2019.

First, the move to linear qualifications removed the opportunity for year 13 students to re-sit AS units at the same time as taking A2 units. Of course, this move to linear was not limited to maths, but maths was unique in that a relatively large proportion of students did re-sit AS units in year 13, and the nature of the subject content is such that students are likely to do much better in AS units after an additional year's teaching.

Year 13 students taking the reformed A level maths had no such opportunity to re-sit part of the qualification. To some extent, therefore, the differences in the grade boundaries reflect this structural change.

Second, it does appear that the 2019 papers were, in general, more demanding. We had some feedback to that effect in the summer, but our investigation has allowed us to estimate the scale of the differences in difficulty. It is very likely that, even if the removal of re-sitting opportunities had not had such an impact, grade boundaries (for at least 3 of the 4 specifications) would have been lower in 2019 to take account of the increased difficulty of the 2019 papers.

The combination of the removal of re-sitting and more difficult papers in 2019 therefore meant that grade boundaries were, in general, lower in 2019 than in 2018 to take account of the increased level of difficulty and also to compensate for the structural changes in the move to linear A levels.

We have considered with the exam boards whether there is a case to re-visit the grade boundaries set in reformed A level maths specifications in 2018. On balance, we believe there is not. Making any changes (to revise grade boundaries downwards) would introduce an unfair advantage for year 12 students taking reformed A level in maths compared to those year 12 students in 2018 who took the legacy specifications (around 1,200 students). We believe that it is right to prioritise comparability of standards for a cohort within a year. We have therefore not asked any of the exam boards to revisit their 2018 awards.

Statistical predictions for AS in 2019

As in previous years, exam boards used statistical predictions, based on students' prior attainment at GCSE, to guide AS awards in 2019. One of the first steps in generating these predictions is to calculate an average GCSE grade for every student in the target age group. In this case, exam boards were using 17-year-old (year 12) students who took their GCSEs in summer 2018. They included all students matched to their prior attainment who had taken 3 or more GCSEs. GCSEs in summer 2018 were a mixed of 9 to 1 and A* to G grades, and exam boards used a mechanism to convert them to a common scale.

However, due to an error, they did not include ungraded results in 9 to 1 GCSEs. This means that just under 4000 students (out of a total of nearly 600,000) were not included in the work to calculate an average GCSE grade because excluding any ungraded results meant that they did not have at least 3 GCSE results. Since exam boards work collaboratively to calculate average GCSE grade, this error affected all boards.

When the error came to light in September, exam boards contacted us and we discussed the implications of the error with them. The impact on the predictions for summer 2019 AS awards varied between boards and subjects, but generally the original predictions used in the awards were slightly generous at grade A and slightly severe at grade E. However, the changes were very small and in many cases, using the amended predictions would not have changed the boundary mark that was presented to senior examiners.

It is also the case that declining entries in the reformed AS subjects have meant predictions are less reliable than in pre-reform AS awards. Exam boards and their senior examiners were aware of this and so they placed greater weight on the judgements of those senior examiners, based on their scrutiny of scripts

In July, senior examiners were content with the boundary marks set in the awards. Exam boards did not believe, therefore, that there was sufficient evidence to re-open the summer 2019 awards, and we agreed with their position. Since results had already been issued, we would have needed to be persuaded that those results already issued were wrong. Given the known unreliability of the predictions for AS, and the fact that senior examiners were content with the awards in the summer, there was no evidence to suggest that this was the case.

We are monitoring the work that exam boards are doing to improve their processes for checking the complex technical work required to generate predictions, to avoid a recurrence of this issue or related issues.

First awards of Pearson Level 1/2 Tech awards

In August Pearson contacted schools and colleges to let them know about a change to the grade points thresholds for the first awards of new BTEC Level 1/2 awards. Pearson made this change in order to maintain standards. We issued a statement¹¹ to make clear that we thought the action taken by Pearson was appropriate, although it was unfortunate that the change was not communicated sooner.

Following the initial announcement, Pearson subsequently announced that it would calculate estimated results for the externally-assessed units for students who had chosen not to re-sit, because they were not aware that the grade points thresholds would change. In our view, this was appropriate to take account, as far as possible, of the likelihood that some students who did not re-sit in summer 2019 might have done so if they had been aware of the revised grade points thresholds.

Having reviewed the issue, we have not changed our view that Pearson's decision was appropriate in order to maintain standards over time and with similar qualifications also offered by Pearson in summer 2019. However, it would have been preferable if this possibility had been made clear much earlier. Pearson told us that one of the key lessons they have learned from the summer was the importance of the timing of any communications. Pearson did not want any communications relating to these Level 1/2 qualifications to be confused with the Level 3 results issue earlier that week and hence did not send emails to affected schools and colleges until Friday 16 August. It is likely that this timing caused some anxiety in those schools and colleges, which was not the intention.

At the time, we said we would reflect in the autumn on the lessons to be learned from this issue. We wrote to Pearson in September to ask them to carry out their own review. It is important that awarding organisations are able to take action to maintain standards, but that should be balanced with the expectations of schools, colleges, students and others. To make this clear, we wrote to all awarding

¹¹ <https://www.gov.uk/government/news/response-to-pearson-changes-to-btec-grading-criteria-2019>

organisations to ask them to review the wording on their websites and/or in their specifications, and to make sure that any statements about the use of grade points thresholds made clear the possibility that these might be changed if there was evidence to support that.

Pearson and some other awarding organisations offering similar qualifications already make use of prior attainment-based predictions. However, there is currently no mechanism to make sure that all awarding organisations are using predictions consistently, and there is no evidence about whether or not grade standards are aligned across awarding organisations within a subject. Ahead of the summer 2020 awards, we will consider whether to extend the approach we have used in level 3 Applied Generals (detailed above) to Level 1/2 qualifications. We plan to explore the use of a common prediction matrix for all awarding organisations in a subject, as this is the most effective way to align standards across awarding organisations.

Conclusion and next steps

In general we were content with the way in which grade boundaries were set in summer 2019. There were a small number of specific issues that we have investigated. In some of those cases there are some lessons to be learned for us and the awarding organisations.

Looking ahead to 2020, we will continue with our work to strengthen the use of senior examiner judgement in setting grade boundaries, so that genuine changes in student performance over time can be reflected in results.

We will also continue our work to monitor the alignment of standards between awarding organisations in large-entry national vocational qualifications taken in schools and colleges.

Appendix A: The use of statistics in awarding

Exam boards have always used statistics, alongside senior examiner judgement, to guide their decisions about where to set grade boundaries in GCSE, AS and A level. Over time, these statistics have become more sophisticated as more data has become available. But there has never been a time when statistics were not used.

Statistics are not the only source of evidence though. Decisions about grade boundaries are based on a number of different sources of evidence – including current and past student work, data about the prior attainment of this year's students compared with previous years, and recommendations from senior examiners.

Statistics used currently take the form of predictions that indicate the proportion of students expected to achieve a certain grade, for example grade 4 and above. Predictions are based on the relationship between prior attainment and national results in a reference year. Exam boards use prior attainment at key stage 2 to predict GCSE outcomes, and prior attainment at GCSE to predict AS and A level outcomes.

Predictions provide a common basis for all exam boards to use and so give us a way to compare grade standards between boards. Each board's prediction is based on the same national results but reflects the prior attainment profile of that board's students.

There are several advantages in using predictions to guide awarding decisions.

- using statistical predictions in a consistent way gives us the best tool to make sure that grade standards between different exam boards are aligned
- when qualifications change, it can be more difficult for senior examiners to judge the quality of student work. In general, students in the first cohorts are likely to perform less well than their predecessors¹², because teachers are less familiar with the new qualifications and there are fewer past papers and other support materials available

Predictions are most reliable when they are based on large numbers of students. With smaller numbers of students, they are less reliable. We and the exam boards take that into account. Where the entry numbers are relatively small, exam boards will balance the use of statistics with the judgements of their senior examiners.

¹² [Investigation into the sawtooth effect in GCSEs, AS and A levels](#)

We have previously commissioned research to evaluate the effectiveness of the prediction approach used for GCSEs¹³, AS and A levels¹⁴. The research supports the approach taken.

¹³ Benton, T., and Sutch, T. (2014) [Analysis of use of Key Stage 2 data in GCSE predictions](#). Coventry: Ofqual. Available at:

¹⁴ Benton, T., and Lin, Y. (2011) [Investigating the relationship between A level results and prior attainment at GCSE](#). Coventry: Ofqual.



© Crown Copyright 2019

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated.

To view this licence, visit

www.nationalarchives.gov.uk/doc/open-government-licence/

or write to

Information Policy Team, The National Archives, Kew, London TW9 4DU

Published by:



Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

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