Research and Analysis

Standardisation of grades in general qualifications in summer 2020: outliers

Identifying students for whom the standardisation model would be unreliable

ofqual

Contents

Executive summary

The circumstances surrounding the award of grades in summer 2020 were unprecedented. In response to the cancellation of assessments in GCSE, AS, A level, Extended Project Qualification and Advanced Extension Awards, we put in place arrangements to collect teacher estimates of the grades that students were most likely to have achieved. This information was provided by schools and colleges in the form of centre assessment grades (CAGs), alongside a rank order of students in each subject, that provided information about their relative expected performance.

The consistency with which this information was generated, and the absolute accuracy of the CAGs submitted, could not be guaranteed. To address any inconsistencies and the likely optimism in the CAGs, we implemented a statistical model to standardise grades across schools and colleges. This was intended to address any advantage or disadvantage to students across the country while also ensuring that national outcomes were broadly maintained. This approach was in line with a direction to Ofqual from the Secretary of State for Education.

Following the issue of standardised A level results, it became apparent that the grades issued did not command public confidence and a great deal of distress was experienced by students, their families, teachers and the wider public. For this we are very sorry. In light of this anguish, we decided to award grades to students that were either the CAGs or the standardised calculated grade, whichever grade was higher.

As with any statistical model, standardisation made assumptions about groups of students. Throughout the development of the standardisation model, ‘outlying students’ were of concern. The term ‘outlier’ is a statistical one. It is used to refer to those students who may be in some way atypical within their centre. For example, they might have a prior attainment profile that makes them quite unique in their centre. This uniqueness might mean that their calculated grade is unreliable. These unusual students were a focus of considerable concern and analysis. We are publishing our analyses in the interests of transparency and so that other researchers can build upon this work.

The characterisation of outlying students in this context differs from the statistical definition that is commonly used. Conventional definitions of outliers refer to observed results in a dataset that are seemingly anomalous due to their distinctiveness from the rest of group. This was not necessarily the case here. Using this conventional definition in relation to standardisation would identify students for who there was a notable difference between the CAG and calculated grade. While there was some public concern relating to such cases, the most notable interest in the lead up to the issuing of A level results was in students who may be atypical in terms of their ability compared to the current and/or historical cohort within their centre. In particular, extremely able students attending centres with more average intakes. In these instances, it is not the **magnitude** of the difference that would have been deemed to be notable (as would be the case with the conventional statistical definition), but the presence of **any** difference between the CAG and the grade actually awarded. It is outlying students of this type who are of particular interest here.

The analyses discussed in this report were considered prior to the release of A level results and sought to identify students who were outliers in individual subjects within their centre on the basis of:

1) their prior attainment, or
2) their CAG

To support more effective identification of outlying students, we applied additional criteria. These included requirements for the student to be at, or close to, the top of the rank order[1] and for the student's calculated grade to be different from their CAG. Additional optional criteria were also employed - the overall generosity of the centre's CAGs for the subject and the uniqueness of the CAG being submitted for the student.

Work took place prior to the release of A level results with a view to considering whether outlying students in receipt of unreliable grades could be identified and new grades estimated. The analyses presented here focus on the A level results issued to students on 13 August 2020, but the issues identified generalise to other qualifications using this approach (such as GCSE).

Analyses based on prior attainment identified 0.4% of entries as potentially outlying. The equivalent analyses based on CAGs identified 0.3% of entries as potentially outlying. However, interpreting these figures is challenging. The grade which students would have achieved had exams not been cancelled cannot be known. It is therefore impossible to know whether the outlying students identified, did indeed receive unreliable calculated grades. The best available way of validating the criteria used to identify outlying students is inspection of individual cases to evaluate the plausibility of the calculated grades. Doing this demonstrates significant uncertainty in whether the student entries identified as outlying, have indeed been disadvantaged through the standardisation process.

[1] The issues considered in this report are largely focused on students who may be considered to be outliers at the top of the ability range and may, therefore, have been disadvantaged. Similar issues apply at the bottom of the ability range. These would, however, mirror the issues at the top of the distribution leading to potential advantage to students through the process. As such, these issues were the subject of less public concern and are therefore not the focus here.

There are 2 other issues in identifying outlying students. First, the criteria used, necessarily, set thresholds against which to evaluate students' characteristics and apply filters which are largely arbitrary. It is not possible to determine whether the threshold values or the criteria design decisions are correct. Second, there is insufficient confidence in the sub-sets of students that are identified using a priori measures to provide an objective determination of whether a student has been disadvantaged.

Regrettably, these limitations meant it was impossible to identify outlier students with unreliable grades in advance of the issue of results. A post-results appeal process was necessary to determine whether a student had received an unreliable grade and to determine the most appropriate replacement grade. Regulatory arrangements were put in place to facilitate such appeals. These would allow consideration of the kinds of technical evidence outlined in this report in conjunction with richer context-specific evidence relating to the individual student and their centre.

We are publishing this work to seek to identify outlying students for who the standardisation process could not be relied upon, in the interests of transparency and so that other researchers can build upon this work. Should any form of statistical standardisation of grades be used in the future, there are lessons to be learned about how best to accommodate unusual students and how best to build confidence in the process.

1. Background

1.1 Context

On 18 March 2020 the Secretary of State for Education told Parliament that, in response to the coronavirus (COVID-19) pandemic, schools and colleges in England would shut to all but the children of critical workers and vulnerable children after 20 March. In line with these measures, exams scheduled for the summer would not take place.

On 23 March 2020, in a written statement to the House of Commons, the Secretary of State explained the government's intention that 'a grade will be awarded this summer based on the best available evidence'. In the direction we received on 31 March 2020, it was confirmed that '[i]n order to mitigate the risk to standards as far as possible, the approach should be standardised across centres' and that distribution of grades should follow a similar profile to that in previous years.

To support this process, centres submitted to exam boards the grades they expected their students to have achieved had exams gone ahead (CAGs) and their judgement of the rank order of students based on their relative abilities in each subject.

Such an approach necessarily brought with it the challenge of consistency in standard applied by centres when providing their CAGs. Formally standardising all teachers across all centres in advance (for example, via national training) to ensure the generation of CAGs was performed in a consistent and equitable way would have been challenging in any circumstance. The magnitude of the task and the context within which it would have been necessary for it to be delivered prevented such an approach.

There is evidence of inconsistency in the accuracy of estimates provided by centres in other contexts (for example, the prediction of grades for the purposes of university admissions). The research literature also identifies differential accuracy across centres with different demographics. Recognising that centres sought to provide holistic judgements in good faith, and subject to quality assurance processes, post-hoc standardisation was seen as an important tool for achieving intra-year fairness to students.

To deliver inter-year fairness, in line with the Secretary of State's direction to maintain the distribution of grades, standardisation also needed to include steps to address any overall generosity or severity observed across the CAGs. As supported by evidence in the research literature[2], it was anticipated that there was likely to be a tendency towards generosity in the CAGs and, therefore, standardisation would also need to ensure that this overall effect was addressed in the production of calculated grades.

To enable the required standardisation of centres' grades, we developed a model to determine the distribution of grades to be awarded to each centre in each subject. The output from the statistical model was then combined with the judgements of ranking from the centre to determine individual student's grades. The development process and the final approach that was applied is documented in detail in our interim report published on A level results day 2020 and is briefly summarised in Section 1.2 below.

Following the issuing of A level results, however, it became clear that the approach we had adopted had failed to command public confidence and had caused significant anguish on the part of students, their families, teachers and the general public. For this we are sorry. We therefore instructed awarding organisations to reissue the A level results, awarding students the higher of their CAG and their calculated grade. On GCSE results day, students received grades on this same revised basis.

Part of the public dissatisfaction with the calculated grades issued to students related to how 'outlying' students might have been treated by the standardisation process. Indeed, in the lead up to results much attention was paid to outlying students. Public discussion was varied, but often focused on students entering for a qualification through a centre where they were atypical in their ability compared to other students in the current year or those who have gone previously.

These concerns were well founded as any statistical approach to predicting the grades of individual students will have limitations. These limitations lead to uncertainty over the calculated grades for individual students (see the discussions of predictive accuracy presented in our interim report; Section 7.6, pp76-81). Further, any statistical model reliant on assumptions about the continuity of results at centre-level would struggle to produce reliable grades for these unusual students.

[2] Dhillon, D (2005) Teachers' estimates of candidates' grades: Curriculum 2000 Advanced Level Qualifications. British Educational Research Journal 31(1) 69-88.

Gill, T (2019) Methods used by teachers to predict final A Level grades for their students. Research Matters. 28 33-42. Cambridge Assessment.

Gill, T and Benton, T (2015) The accuracy of forecast grades for OCR A levels in June 2014. Cambridge, UK: Cambridge Assessment.

Gill, T and Benton, T (2013) The accuracy of forecast grades for OCR A levels in June 2012. Cambridge, UK: Cambridge Assessment.

The subject of this report is a technical consideration of the detection of these outlying students; how they might be defined in the context of standardisation; attempts to detect their presence and issues related to the calculation of a more reliable grade once they are identified. The issues discussed are relevant to the standardisation of all qualifications types for which we put in place explicit regulations to standardise results in summer 2020[3]. However, for the purposes of simplicity, the predominant focus in this report is on A levels.

1.2 The standardisation model

At the highest level, there were 3 approaches to standardisation that were considered for summer 2020. These were:

1) **Macro-level standardisation** where the adjustment applied is defined by a population-level relationship that is applied to the whole cohort in a subject.
2) **Meso-level standardisation** where centre-level statistical estimates are used to standardise each centre in each subject.
3) **Micro-level standardisation** where estimates are formed based on the characteristics of individual students.

A full consideration of these approaches is provided in Section 6.1 of our interim report. The latter 2 of these approaches is relevant to the issues considered here.

The approach implemented in summer 2020 was meso-level standardisation. This meant that statistical analyses were performed at the centre-level with the aim of achieving fairness between centres. The results for individual students were then determined by the information provided by teachers within the constraints of the statistical prediction for the centre in each subject.

The full details of the standardisation model that exam boards were required to implement to determine students' calculated grades are provided in Section 8 of our interim report and are codified in our regulatory requirements.

In brief, the standardisation model sought to predict the distribution of grades for each centre in each subject based on 3 key pieces of data:

A. The distribution of grades achieved by each centre in that subject over recent years. The number of years across which historical results were aggregated varied by qualification type[4].

[3] The qualification types included were GCSE, AS, A level, Extended Project Qualifications and Advanced Extension Awards

[4] For AS and A level, 3 years of historical data were used. For GCSE, two years of data were used for reformed specifications that were first awarded in 2017 and 2018 and a single year was used for those first awarded in 2019. See Section 7.2 of the interim report. The handling of centres without historical performance data is considered in Section 8.4.1 of the interim report.

B. The prior attainment of the cohort of students making up the historical grade distributions in each centre for each subject. For AS and A level qualifications, students' prior attainment is defined as their mean GCSE performance and for GCSE it is based on their KS2 results.

C. The prior attainment of the cohort of students entering for the subject with each centre in summer 2020 (following the same definition of prior attainment as specified in B).

The basis of the approach was to use the historical grade distribution for each centre in each subject (A) as a start point. This distribution was then adjusted based on the difference in prior attainment profiles of the cohorts within the centre – those from previous years (B) and those from the current year (C).

Having established the predicted grade distribution, individual students were awarded grades based on the rank order as submitted by the centre, meeting the predicted grade distribution as closely as possible. An overview of this process is provided in slide 18 of our published summer symposium materials and described in the accompanying video.

As meso-level approaches such as these apply statistical models at the centre-level they necessarily make statistical assumptions at the centre-level too. In the case of the model outlined above, these assumptions relate to the continuity of results from previous years and the rate at which differences in the prior attainment profile of students over time should affect the outcomes. An alternative approach would be to operate at the micro-level. This would mean setting aside the contextual information provided by the centre that students attend and relying solely on the statistical indicators relating to the individual, such as their individual, rather than group, prior attainment.

The use of measures of prior attainment are commonplace in research studies seeking to control for differences in the underlying ability of students/participants,[5] and are also routinely applied for operational predictive purposes by exam boards.

These analyses, however, tend not to rely on the prior attainment at the individual level due to the limited predictive accuracy for individual students. An individual student with a high prior attainment may be more **likely** to achieve a higher grade in a subject compared with a student with a lower prior attainment, however, it would be inappropriate for this to **predetermine** the results for individuals on this basis on a student-by-student level.

[5] Pinot de Moira, A., Meadows, M.L. & Baird, J-A. (2019) The SES equity gap and the reform from modular to linear GCSE mathematics, *British Educational Research Journal*.

Typical correlations between prior attainment and attainment in a GCSE are 0.34 to 0.76[6] and are 0.57 to 0.71 at A level[7]. While acting as a valuable indicator when applied to groups, such as in meso-standardisation approaches, it would be an over-interpretation of the measures to use them to predict the outcome for an individual student.

It is also unclear what the role would be for the rank order information provided by centres if such an approach was used. A key motivator behind collecting the rank order judgements is that they provided a measure of relative ability that cannot be sufficiently captured by purely statistical means. Putting greater reliance on statistical measures relating to an individual student would, therefore, risk disordering the rank order submitted by the centre in an indefensible, and likely invalid, way. Taking such an approach that would override teachers' rank ordering of students would reduce the weight of the centre judgements in the process.

These issues relating to micro-level standardisation led to it being discounted as the approach to take, however, they are central to the consideration of outlying students in terms of both their detection and potential remedy. To identify a student through statistical means for whom it might be deemed to rely on the group relationship or for whom the outcome from the process is objectively unfair would need to put greater reliance on the student-level statistical evidence. The challenges of doing so are fundamental to the issues discussed here.

The statistical models that underpin meso-level standardisation approaches are, of course, not without their limitations. By definition, these statistical models rely on group level statistics (subjects within centres) and assume all students are part of the same statistical relationship. These approaches are limited in their ability to accommodate students that are atypical compared to the rest of the group, and, therefore, where the statistical model does not hold. There are 2 challenges that, therefore, exist when seeking to address these limitations. The first is the identification of the atypical subset of students who may have legitimate claims as to the inappropriateness of the model being used to estimate their grade, the second is how to award grades to these students if/when they can be successfully identified. These challenges are the subject of Sections 3, 4 and 5. In the next section, consideration is given to the definition of 'outliers' in the context of standardisation.

[6] Benton, T. and Sutch, T. (2013). Exploring the value of GCSE prediction matrices based upon attainment at Key Stage 2. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment

[7] Benton, T. and Bramley, T. (2017). Some thoughts on the 'Comparative Progression Analysis' method for investigating inter-subject comparability. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

2. Definition of outliers

Classically in statistical analyses, outliers are identified as observed data points that differ notably from other observations in the dataset and where the other data follow some underlying relationship. A simple example is shown in Figure 1 for a fictitious dataset with an outlier shown in red (above the main cluster of points).

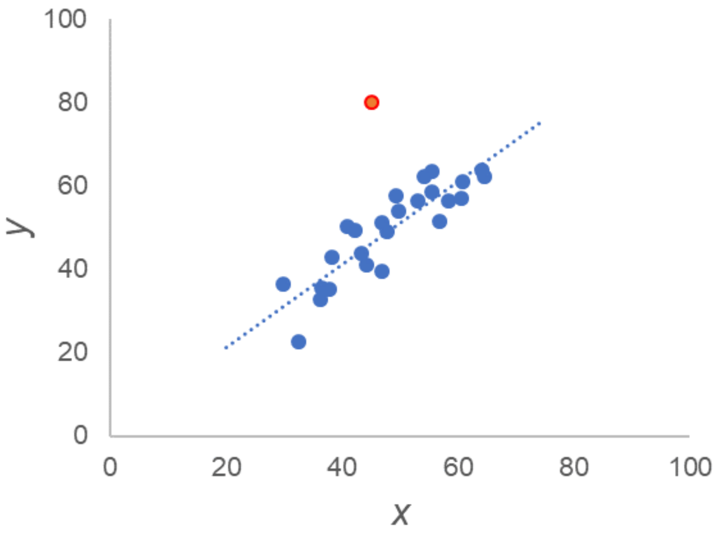

Figure 1 General representation of an outlying data point

Outlying data points can exist for a number of reasons.

Scenario 1:

Most simply, the data point may be erroneous – it has arisen due to some form of mistake or measurement error that has occurred through the process of its collection.

Scenario 2:

Alternatively, there may be some feature of the data subject that means it does not obey the underlying relationship or model associated with the other data points. There may be some confounding feature related to that data point or some unmodelled characteristic that affects the data point more than others meaning that it does not follow the apparent relationship followed by the rest of the data.

Scenario 3:

Finally, an outlier point may be legitimate and a faithful representation of the relationship that underpins the other data points but arises simply due to chance. Summary representations of relationships such as the dotted line shown in the Figure 1 or summary statistics (such as mean and standard deviation) are simplifications of the underlying probability distributions which reflect the natural variation of the measures. These summaries distil down information contained in the many data points down to a single or small number of more interpretable measures that summarise those data. The underlying distributions may have long ‘tails’ meaning that data points relatively far from the line that summaries the relationship are probabilistically possible but are very unlikely to have occurred given the size of the dataset, despite being formed from the same underlying distributions.

To provide a basis on which to consider outlying students in the context of the standardisation model, it is helpful to first consider an analogous relationship in a typical exam year. For example, a relationship such as that shown in Figure 1 could be the relationship between students’ prior attainment and the marks they achieved on an assessment. Prior attainment is a covariate or measure commonly used in both operational and research activities to control for, or to explain, variations in the performance of students within a population. For operational standard setting purposes, exam boards use the mean GCSE results as an indicator of the overall ability of a group of students when performing analyses relevant to AS and A level qualifications. For GCSE, the equivalent prior attainment measure is provided by students’ KS2 test results.

Outliers that arise through Scenario 1 – instances of an error or mistake occurring in the process – are easy to conceptualise in relation to a typical exam series. This could arise from an administrative mistake or instance of objectively errant marking that occurs through the operational processes that support delivery of the assessment. These scenarios could lead to a student’s mark being incorrectly recorded and/or grade being awarded. Existing arrangements such as reviews of marking and moderation which include administrative checks are required to be put in place by exam boards to remedy such instances.

In Scenario 2 – that where the relationship characterising the data for most students is inappropriate for an individual – would have no consequence for individual students in a typical year. This is because the student’s grade is determined purely by the number of marks they achieve relative to the grade boundaries, irrespective of the extent to which they obey the underlying relationship. These cases where other covariates may indicate differential performance may be interesting for research purposes (for example, if exploring of student characteristics or experiences that might predict educational outcomes) but have no operational consequences.

The third scenario – where a student’s result is statistically unlikely, but theoretically possible based on the underlying relationship – would again have no consequence for the individual in a typical series but may arise for 2 distinct reasons. The first (Scenario 3a) may occur simply because the student has progressed to a significantly greater or lesser extent than is typical for a student with that level of prior attainment. This could be for a range of localised environmental or developmental reasons. Alternatively (Scenario 3b), a student may be at an extreme of the distribution because of a surprising one-off performance due to inherent uncertainty in the assessment process (for example, due to the sample of questions included in a particular assessment being particularly aligned or misaligned to the student's strengths).

2.1 The characteristics of 'outlying' students

It is helpful to reflect on why students might be statistically speaking 'outliers' in the context of the standardisation model. So far, we have focused on outlying **observations**. This follows the conventional statistical definition and highlights an important difference – and added challenge – to the process of considering outlying students in relation to the standardisation process.

Before results were issued, concerns were not based on evidence of students being observed as outlying on the outcome measure (that being the difference between CAG and calculated grade). Rather, concern was for students who might be of atypical ability for their centre or who might be unusual in terms of their input characteristics which might lead to small, but personally highly significant, differences between CAGs and calculated grades (for example, A* versus A). Given the stakes associated with grades these differences were understandably considered unfair.

In essence, these are concerns that the atypical nature of the student (given the characteristics of other students attending their centre), would mean that model and its assumptions were inappropriate for them (scenario 2), or that they had sustainably and predictably bucked the trend (scenario 3a) but that the full extent of this was not visible in the calculated grades.

The challenge is therefore not to **detect** outlying students on the basis of observations (the difference between the calculated grade and CAG) as is usually the case, but to **predict** the future presence of an 'incorrect' observation due to outlying input characteristics. To heighten the challenge, these outlying input characteristics may or may not be observable as a large difference between the CAG and calculated grade or as a large difference between the potential outlying student's grade and other students in the centre.

To explore this further, the example considered in the introduction is revisited - an instance where the prior attainment of a student may be atypically high compared to other students in the current cohort and/or those that have gone before in previous cohorts at the centre. There are a range of possible scenarios that might arise which highlight the challenges of detecting outlying students before the release of results, Which of these scenarios occurs in each case is, however, unknown and unknowable:

A. The centre believes the student would have been outlying in terms of his/her performance compared to others in the current cohort and those that have gone previously. This is in-line with his/her higher than typical prior attainment. This is reflected in a CAG that is higher than is precedented for the centre. The assumptions inherent in the statistical model are **inappropriate** and unfortunately the student receives a calculated grade lower than the CAG, and lower than would have been the case had exams not been cancelled.

B. The centre believes the student would have been outlying in terms of his/her performance compared to others in the current cohort and those that have gone previously. This is in-line with his/her higher than typical prior attainment. This is reflected this in a CAG that is higher than is precedented for the centre. The assumptions inherent in the statistical model, however, are **appropriate** and the student receives a calculated grade that matches the grade they would have achieved had exams not been cancelled but is lower than the CAG. The student and centre are understandably disappointed but if exams had gone ahead the student would not have actually achieved the CAG.

C. The centre believes the student would have performed well, but not exceptionally so, and not to the extent indicated by his/her atypical prior attainment. This is reflected by a high position in the rank order, but a CAG that is not unprecedented within the centre. The assumptions inherent in the statistical model are **appropriate** and the student receives a calculated grade that matches the CAG and the grade they would have achieved had exams not been cancelled. The centre's expectations have been met and the student has been fairly awarded.

D. The centre believes the student would have performed well, but not exceptionally so, and not to the extent indicated by his/her atypical prior attainment. This is reflected by a high position in the rank order, but a CAG that is not unprecedented within the centre. The assumptions inherent in the statistical model are **inappropriate** and the student receives a calculated grade that matches the CAG but is unfortunately lower than the grade they would have achieved had exams not been cancelled. The centre's expectations have been met, but the student has been under-rewarded.

As can be seen, each of these scenarios leads to a different conclusion, however, all of the evidence (statistical and judgemental) feeding into scenarios A and B are identical and unfortunately cannot be separated. The same is true of the evidence feeding into scenarios C and D. Indeed, the only distinction that can be made between the evidence available across all four scenarios are the expectations of the centre articulated through the CAGs – the statistical evidence and rank order is identical in all cases. As discussed previously, the inconsistency across centres in the approach taken to generating CAGs and the potential unfair advantage or disadvantage that might result means it would be difficult to rely on such evidence in absolute terms.

This leads to another factor that increases the challenge of reliably separating the 4 scenarios above – the tendency for the centre to be seemingly generous, accurate or lenient in their CAGs overall. For example, if the centre's CAGs appear to be accurate or severe for the majority of the cohort, this suggests that, between scenarios A and B, scenario A is the more likely. However, if there is a tendency for the centre to be generous overall, scenario B may be the more likely. In an instance where an overall downward adjustment from CAGs to calculated grades is indicated, it is difficult to judge whether an outlying student is an outlier to such an extent that their CAG should nonetheless stand.

In summary, the primary challenge is to reliably identify students who are outlying in terms of their input characteristics and where this would lead to an incorrect adjustment to their CAG. This distinction needs to be made based on a combination of statistical evidence which does not clearly separate atypical and typical students and absolute judgemental evidence from centres which contains known inconsistencies.

2.2 Outlying outcomes

Following the issue of results, there was understandably a great deal of concern about students with calculated grades which differed notably from their CAGs. These differences can be thought of as the second, more conventional type of outlier – one based on output measures. Again, to contextualise this issue, it is helpful to consider it in relation to normal operation.

In a normal year there is uncertainty about whether a student will achieve a particular grade; were this not the case, there would be less of a role for formal assessment. In the vast majority of cases, students' grades tend to be within 1 grade of that which their teachers anticipated. This pattern is seen in the relationship between predicted and actual A level grades. Indeed, there is strong evidence that the estimates provided by teachers tend to be generous[8]. It is important, however, to note not just

[8] Dhillon, D (2005) Teachers' estimates of candidates' grades: Curriculum 2000 Advanced Level Qualifications. British Educational Research Journal 31(1) 69-88.

Gill, T and Benton, T (2015). The accuracy of forecast grades for OCR A levels in June 2014. Research Matters. Cambridge, UK: Cambridge Assessment.

Delap, M (1994) An investigation into the accuracy of A-level predicted grades. Educational Research 36(2) 135-148

Delap, M (1995) Teachers' Estimates of Candidates' Performances in Public Examinations. Assessment in Education: Principles, Policy & Practice. 2(1) 75-92.

the accuracy of these estimates, but also the distribution of the inaccuracy in the estimates. It would seem intuitive that any inaccuracy in estimates would be, in the vast majority of cases, just by a single grade. It is also accepted that students may have a good or bad day meaning that their exam performance is slightly better or worse than expected and so they may achieve a grade higher or lower than predicted[9].

However, following the issue of results, differences of this magnitude between calculated grades and CAGs were not viewed as benign. Rather, these differences had a notable and negative impact on the public's acceptance of the calculated grades. With hindsight it is clear that this was because students had not had the opportunity to demonstrate what they knew and could do – the differences were not explained by assessment evidence and were a product of a standardisation model. The public response to differences of 1 grade between calculated grades and CAGs provides context for our consideration of instances of larger differences.

Replicated below from 2 relevant publications are representations of the distributions of differences between estimated grades (similar in many respects to CAGs) and actual grades. Table 1 shows a cross tabulation of estimated grades against the actual grades awarded to students as reported in Dhillon (2005)[10]. This work considered the accuracy of estimates across A levels offered by AQA in chemistry, English literature, history, mathematics and psychology. While showing the overall bias towards generosity, of note is the significant proportion of students receiving a grade more than 1 grade away from the estimate. It is reasonable to suspect that calculated grades in summer 2020 that deviated from CAGs to this degree would have been considered anomalous by the public.

While deviations from expectation were seen as incorrect or anomalous products of the standardisation process, such differences exist routinely in a normal year. In this example, 8.8% of results are 2 grades or more from expectation. The key difference, of course, is that these instances occur as a result of assessment evidence produced by students and this has a marked impact on their acceptability.

[9] Chamberlain, S., Public perceptions of reliability (Ofqual/10/4708) in *Reliability of assessment: compendium*. Ofqual 2013.

[10] This research was conducted prior to the introduction of grade A* at A level and, therefore, the grade set runs from grades A to U.

Table 1 Number of candidates achieving grades A to U across the selected A levels as a function of estimated grade (the shaded area corresponds to optimistic estimates) Reproduced from Dhillon (2005)[11]

	A Awarded	B Awarded	C Awarded	D Awarded	E Awarded	Ungraded	Totals by estimate
A Estimated	10541	3957	665	78	11	4	15256
B Estimated	2202	6684	4731	1075	167	36	14895
C Estimated	303	2447	6499	4805	1294	256	15604
D Estimated	26	331	1889	3824	2562	675	9307
E Estimated	4	49	294	1268	2157	1378	5150
U Estimated	1	3	10	47	184	367	612
Totals by award	13077	13471	14088	11097	6375	2716	60824

An alternative representation of deviations from expectation is presented by Gill and Benton (2015). This considers the results more explicitly from the perspective of their use in higher education admissions. Table 2 shows the difference between actual results and estimated results in terms of the points scores using the UCAS tariff[12].

As can be seen from these distributions, there is a wide range of differences between the actual point scores achieved compared to those that were expected, with a tendency towards generosity. There are a small number of students with an extremely large difference between expected and actual outcomes. It would seem reasonable to suspect that the most significant outliers in this distribution, such as those where estimates were more than 160 points (or 8 grades on aggregate) away from the estimate, may have occurred as a result of some significant unexpected personal or particularly local event that was not or could not be compensated for through the established special considerations processes. Isolated instances such as this are very unlikely to be predictable. More moderate deviations from the expected results, however, are more likely to be associated with students simply delivering a different level of performance to that which was anticipated; be that due to the limitations of the estimate itself or features of the performance.

[11] Note that the headings of the table have been modified for the purposes of accessibility.

[12] For reference, A* - 140 points, A = 120 points, B = 100 points etc.

Table 2 Distribution of actual difference between final and forecast points score. Reproduced from Gill and Benton (2015)

Actual Difference	Frequency	Percent
-220	1	0.02
-200	3	0.05
-180	4	0.06
-160	18	0.28
-140	27	0.42
-120	48	0.74
-100	168	2.58
-80	302	4.64
-60	609	9.37
-40	1097	16.87
-20	1430	21.99
0	1509	23.21
20	833	12.81
40	337	5.18
60	88	1.35
80	23	0.35
100	4	0.06
120	0	0.00
140	1	0.02

Take, for example, cases where the difference between expected and actual points scored was 100 points. This is a relatively small proportion of students (2.58%) within this sample, but if representative across the A level cohort, equates to several thousand students. For a student taking 3 A levels this might correspond to a single subject being 5 grades away from that predicted (A* to E or A to U) or, more likely, being 2 grades away on 2 subjects and 1 away on another (for example, AAB to BCC or some other combination). Again, in the context of standardisation, such results would likely have been seen as anomalous or lacking credibility, particularly given the absence of any direct evidence from the student to inform the grade awarded.

One consideration during the issuing of calculated grades was the potential impact that the standardisation process might have on students' combinations of grades. For example, were there to have been a notable change in the rate of balanced grade combinations achieved by students (for example, AAA or BCC) in favour of imbalanced, less common, profiles (for example, A*BE or ADU) this may have indicated a prevalence of anomalous individual grades. These analyses were reported in Section 9.6 of the summer 2020 interim report for calculated grades awarded as a result of standardisation and are replicated in Table 3 for convenience.

Table 3 Proportion of students achieving the 20 most common grade combinations (2017 to 2020). 2020 data are based on calculated grades.

	Percentage of total students			
Grade combination	2017	2018	2019	2020
BBC	7.9	7.9	7.8	7.7
BCC	7.0	7.1	7.1	6.9
ABB	7.0	7.0	7.0	6.9
AAB	5.9	6.0	5.9	6.0
BCD	5.4	5.4	5.3	5.4
CCD	4.9	5.0	5.2	4.9
ABC	5.2	5.1	4.9	5.3
BBB	4.6	4.7	4.5	4.3
A*AA	4.2	4.1	3.8	4.1
CDD	3.6	3.7	4.0	3.5
AAA	3.5	3.7	3.6	3.5
CCC	3.2	3.3	3.3	3.3
A*AB	3.2	2.8	2.7	2.9
CDE	2.6	2.5	2.9	2.5
A*A*A	2.8	2.7	2.5	2.9
BBD	2.2	2.2	2.1	2.2
DDE	1.6	1.7	1.9	1.6
ACC	1.6	1.6	1.6	1.7
A*A*A*	1.5	1.5	1.6	1.7
BDD	1.6	1.6	1.5	1.5

This demonstrates that the proportion of students receiving the most common grade combinations were highly comparable in 2020 (based on calculated grades) with previous years and with any differences being within the natural year-on-year variation.

When considering unexpected results, much of the focus was, understandably, on those students where their CAGs were higher than their calculated grades. It is also important, however, to consider the reverse scenario, where the calculated grade was higher than the CAG, and in some instances significantly so. These instances undermined confidence in the standardisation model.

A useful scenario to consider, which was given particular attention through the design of the standardisation process, was the treatment of students who were estimated to be ungraded. This was of particular interest for 2 reasons. First, the stakes around the transition from ungraded to grade E are different to the transition between grades; it is the difference between the student being awarded a qualification and receiving no qualification at all. Second, views were expressed regarding the greater ease that it was felt teachers would have in correctly identifying students who would have failed to achieve a grade had exams taken place. These views are not, however, supported by previous research evidence. As an example, revisiting the data presented in Table 1 shows that, of the 612 students estimated to be ungraded, 40% (245) of those estimates were incorrect with students achieving a grade with some of these differences being considerable. This is replicated in other similar studies.

In summary, large differences between CAGs and calculated grades appeared implausible to the public. This, combined with the lack of agency students had over their results, significantly undermined public confidence.

However, it is clear that, in a typical year, there are a small proportion of students (but large in number when considered across the national cohort) with actual grades notably different from those they are predicted. This is true for individual grades and combinations of grades across a student’s subjects.

From a technical perspective, this means that a student having a notably different CAG from their calculated grade is only a weak indication that the assumptions underlying the standardisation model were inappropriate for that student. In seeking to identify outlying students then this information is not as helpful as one would wish. Detection criteria is the subject of the next section.

3. Detection criteria

A range of statistical techniques are available to detect the presence and identity of outliers in data. These might be based on rudimentary statistical measures such as the location of data point in relation to the inter-quartile range or might involve sophisticated approaches embedded within machine learning solutions.

As described, the definition of outliers and the challenge of identifying them in this context is atypical. An obvious pre-existing solution is, therefore, not available. To attempt to identify students for who the assumptions in the statistical model were not appropriate a pragmatic approach was taken. This involved defining sets of criteria through which students were filtered. These approaches were considered prior to the issuing of results to determine whether they should be incorporated into the standardisation approach. The criteria sets were designed around instances where the assumptions of the model may be broken and are described below.

The design of the standardisation model was such that 2 key assumptions were made. The first relates to the continuity of student attainment for each centre over time. The second relates to the impact that variations in the prior attainment of each centre's cohort would have on their results. One of the limitations is, therefore, that if students deviate significantly from these historical expectations or their behaviour does not follow the expected difference in value-added relationship, they may be unfairly advantaged or disadvantaged.

As highlighted in the previous section, there are 3 types of outlying students considered here:

1. Those who are atypical in terms of their prior attainment and, therefore, the constraints of the centre's statistical prediction may be unfair.
2. Those who are considered outliers in terms of their current ability meaning that fitting them into a smooth distribution along with their peers, as defined by the model, may be unfair.
3. Those who have a notable difference between their CAG and calculated grade.

Sets of criteria to identify outliers based on these categorisations are explored below.

3.1 Distribution of prior attainment

The first set of criteria used measures of students' prior attainment as the primary indicator. The rationale is that an atypical prior attainment measure indicates that a student might be more able than any student that has attended in the past and/or attends in the current year. This may mean the assumption of continued historical outcomes (even when accounting for overall differences in the prior attainment of the centre's cohort) is insufficient to allow for the outlying student.

Following standard operational definitions used for A level, students are considered to be "matched" to prior attainment if they meet the following criteria:

- they have a valid record of GCSE results from 2 years previously that can be identified based on the student's Unique Candidate Identifier, forename, surname, date of birth and/or sex[13]
- they have at least 3 valid GCSE results from 2 years previous
- they are the target age for the qualification – 18 years old for A level – as of 31 August in the year that they complete the qualification.[14]

The need to match students to their prior attainment highlights a limitation of this approach; students without measures of prior attainment cannot be identified as outlying on this basis. This is less of an issue for A level given the high proportion of students who meet the criteria above compared to other qualifications. 83% of entries in summer 2020 were from students who could be matched. However, 125,577 entries were unmatched which corresponded to 56,636 unique students.

Once matched students are identified, the next step is to identify which of them have measures that are atypical within their centre for the subject. The relatively small numbers of students entering for subjects at A level makes the reliable parameterisation of the distribution in the form of summary statistics challenging. A pragmatic approach is taken, assuming a normal distribution of prior attainment within the centre for each subject and calculating a z-score for each matched student.

This is calculated using the standard definition:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where μ_j is the mean of students' prior attainment in centre j, σ_j is the standard deviation and x_{ij} is the measure of prior attainment for student i.

The distribution of z-scores within a centre is continuous. It is, therefore, necessary to identify a threshold above which a data point would be deemed to be an outlier and below which they are not. The value of this threshold is arbitrary, but necessary

[13] See Annex B of the published requirements placed on exam boards to deliver the award of standardised grades for more detailed specification of the matching requirements.

[14] This definition matches that used operationally to produce statistical predictions used routinely for supporting the maintenance of standards. The decision was taken through the design of the standardisation model to follow the same definition.

for the purposes of categorisation or filtering. Based on a standard normal distribution, a commonly used reference point is $Z = 2.0$ (the value 2 standard deviations above the mean value). Assuming a standard normal distribution, this point means that 2.3% of data points would be expected to fall above this threshold, if the underlying data follows this relationship. While this threshold is arbitrary, it is a sufficiently stringent measure to identify the most outlying students while also being sufficiently permissive to overcome the limitation of the necessary distributional assumptions, identifying students whose prior attainment is either at the extremes of the distribution or who should not be considered to be part of the same distribution.

Two variants of the filtering for prior attainment were applied. The first version is built using distributions of prior attainment based only on students entering in summer 2020. The second version built the distributions of prior attainment using data from all the historical years of data included in the standardisation model. In the case of A level, this included 3 years of historical data plus the current year. A discussion of the selection of years of historical data used in the standardisation model is provided in Section 7.2 of the interim report.

Applying the above primary filter does not, in itself, ensure that the students identified are sufficiently atypical for the model to be inappropriate. First, there will always be students at the extremes of the distribution. This does not mean the assumptions of the model have been violated.

Second, a limitation of this approach is that there is not a one-to-one relationship between prior attainment and actual attainment. Even if a student is an extreme case in terms of his/her prior attainment, it does not mean they are an extreme case in terms of their **current** attainment. Were this to be the case, it would have likely been appropriate for the whole standardisation process to be based on a micro-level approach, as discussed in Section 1.2. It is very possible that a student with the highest prior attainment within a centre, is not ranked the highest and/or has not been allocated the highest CAG. Equally, a student with the weakest prior attainment measure may have developed at a faster than average rate, or the circumstances which led to them performing relatively poorly during their GCSEs may have changed. Such a student may well not be given the lowest rank within a centre. To improve the chances of the approach identifying genuine outlying students, and to remove false positives, it is therefore necessary to apply additional criteria to filter the data.

The following additional criteria were, therefore, added to the criterion of having a z-score greater than 2.0:

i. The student must be ranked within the top 2 students in the subject for the centre. This attempts to isolate the students most likely to be outlying at the top of the mark distribution while also accounting for the occurrence of more than 1 atypical student – a student not occurring at the extreme of the rank order could not be rationally defined as being atypical in terms of his/her expected performance.

ii. The student must have a different calculated grade from their CAG. It would be illogical to consider the model inappropriate for a student if it had confirmed the judgement of the centre.

Application of these 2 filters were necessary to provide a meaningful sub-set of entries who may be outliers. Two further optional criteria were also used:

iii. The centre must not have been more generous (based on a comparison of the predicted grade distribution and the distribution of CAGs) than the average level of generosity of all centres. As is explored further below, the apparent generosity, accuracy or severity of the centres CAGs proves to be a complicating factor when confidently identifying instances where a student may have been disadvantaged

iv. The CAG must be atypical among the current cohort. To mirror criterion i, this requirement is that the student must be either the only student, or only one of 2 students receiving the CAG from the centre for that subject. This reflects the atypical nature of these potential outlying students and demonstrates a separation from other students in the centre’s cohort.

3.2 Distribution of CAGs

The second set of criteria use the distribution of CAGs as the primary basis for filtering. This is a legitimate basis – and arguably a stronger basis than prior attainment – for 2 key reasons. First, it is the most recent indicator of the ability of students. This overcomes issues with the indirect nature of the prior attainment measures explored above. Second, judgements from centres are the only direct, subject-specific, indicator that a student may be atypical in terms of their ability relative to others in their immediate cohort. Were the student not distinguishable from those around them on the basis of their CAG, it wouldn’t be logical to claim that they were atypical in terms of their ability relative to their peers.

As an initial basis for the filter, similar to the criteria based on prior attainment, centre-level distributions for each subject were built, this time based on CAGs. This approach has similar limitations to those of the prior attainment-based distributions when parameterising these distributions, with the added issue of the discrete nature of the underlying data. However, pragmatically, a similar approach was applied based on the same distributional assumptions. Students with z-scores on the CAG distribution greater than 2.0 are considered to be outlying on this measure.

The use of CAGs as a basis of identifying outlying students is not without its limitations. The purpose of the standardisation process is to remove the inconsistencies and potential biases CAGs are likely to contain. It is, therefore, somewhat circular to rely on this information as a basis for to identifying anomalies. Issues with the reliability of these data risk the legitimate identification of outlying students. To guard against this, the filter regarding overall generosity of CAGs relative to the calculated grades, as included above, is replicated here. The rationale is that in instances where the CAGs are particularly generous, their credibility is insufficient to consider them appropriate for the purposes of identifying outlying students. This criterion is likely to be more critical here than for the filter based on prior attainment.

The following additional refining criteria were then added:

i. The student must be ranked within the top 2 students in the subject for the centre.

ii. The student must have a different calculated grade from their CAG (mirroring the criteria defined above).

Similar to the prior attainment-based approach, the following 2 optional criteria were also applied:

iii. The centre must not have been more generous (based on a comparison of the predicted grade distribution and the distribution of CAGs) than the average level of generosity.

iv. The CAG must be atypical among the current cohort.

3.3 Difference between CAGs and calculated grade

The final type of outlying student discussed in Section 2 is one whose calculated grade and CAG are notably different. This is the most relevant **outcome** measure to define in this context. The existence of a difference between the CAG and calculated grade has been included in the criteria sets defined above but is not explored further as a mechanism to identify outlying students in isolation. This is for 3 key reasons.

First, for moderately sized cohorts, with anything other than a particularly able or particularly weak historical distribution, it is unlikely that students for who the model assumptions are not appropriate would lead to a particularly large, outlying, difference between the CAG and calculated grade. Considering atypically able students within a cohort, the student would likely be the number 1 rank within the centre. Given this position, even if the model assumptions do not hold, it is not possible for the student to be awarded a grade lower than the number 2 ranked student due to the retention of the centre's rank order through the process. This provides a protection against a particularly large difference in grade. For a typical distribution of students, it is unlikely that this would lead to a difference of more than 1 or 2 grades between the CAG and calculated grade. While we recognise the potential personal impact of such a difference between the expectations and awarded grade for the students, statistically speaking the difference is not numerically large enough to render the outcome measure an outlier.

On a related point, large differences between CAGs and calculated grades are more likely to occur lower down in the grade distribution and not due to outlying students at the top of the ability range. For example, they occur where there is an overall view, reflected in the distribution of CAGs, that the cohort for a particular centre would perform far better than historical performance would suggest (to which calculated grades for were anchored) to the point where such outcomes would be highly statistically surprising. A real example of this can be seen below. In these cases, the distributions of CAGs submitted by the centre are significantly out-of-line with the grades achieved in the subject over recent years. This would give rise to students with CAGs of grades A in the case of biology and physics and grades A and B in the case of chemistry, being awarded significantly lower grades than the CAGs. Particular differences of note are emboldened, and all figures are cumulative percentages showing the percentage of students at the quoted grade or higher.

Biology	A*	A	B	C	D	E
Historical outcomes (2017-19)	19.1%	**48.5%**	**69.1%**	85.3%	95.6%	97.1%
CAGs 2020	35.3%	**100.0%**	**100.0%**	100.0%	100.0%	100.0%

Chemistry	A*	A	B	C	D	E
Historical outcomes (2017-19)	**34.2%**	**50.0%**	**67.1%**	84.2%	91.5%	96.3%
CAGs 2020	**75.6%**	**97.6%**	**100.0%**	100.0%	100.0%	100.0%

Physics	A*	A	B	C	D	E
Historical outcomes (2017-19)	**22.6%**	**51.6%**	**64.5%**	80.6%	93.5%	96.8%
CAGs 2020	**43.3%**	**100.0%**	**100.0%**	100.0%	100.0%	100.0%

In such cases, scrutinising the standardisation model to identify the source of an anomaly leading to the differences between the calculated grades and CAGs will likely be fruitless. There may have been changes within the centre to explain the step-change in performance and it is not possible to use statistics to separate these cases from those where such a notable change in outcomes is not credible.

Finally, as described in Section 2.2, instances of large differences between estimated grades and actual grades are not uncommon, albeit the centre and/or student may be surprised by the result. The existence of a large difference does not, therefore, in isolation suggest a student has been unfairly treated in the standardisation process. Given this, the more useful role of this outcome measure has been incorporated into the criteria sets described above.

4. Results and case studies

Interpreting the results of the filters described in the previous section is challenging and highlights a fundamental issue. It is not possible to objectively evaluate whether or not the students identified by the criteria have indeed been unfairly treated through the standardisation process. It will never be possible to establish the appropriateness of the individual grades – whether the ‘correct’ grade for these students was the CAG, the calculated grade or some other grade. This is due to the counter-factual nature of the issue.

To interpret the results it is, therefore, necessary to either inspect the cases that are identified and consider, judgementally rather than statistically, whether the scenario appears plausible or to use secondary indications of the effectiveness of the criteria.

An opportunity to validate whether or not students had been disadvantaged through the standardisation process would have been to compare the outputs with the entries received and grades achieved in the exceptional autumn series. This is no longer appropriate, however, for 2 reasons. First, the size of the A level entries for the majority of A levels in autumn 2020 are significantly smaller than was anticipated when the series was conceived meaning the reliability of such a comparison would be low. This is due to the reduced need for the series due to the ultimate award of CAGs (where higher than the calculated grade) in the summer series. Secondly, the grade boundaries being set are seeking to be at a level equivalent to the performances what would have been required to realise the outcomes in the summer. These will not be comparable with previous years (which was the basis of the standardisation approach) due to their generosity meaning these results do not provide a meaningful comparison with those produced through the standardisation process.

A key piece of information is also missing from the process – the views of the centre as to whether or not the student is indeed atypical. While the challenges of the objectivity and reliability of these judgements remain, it may be valuable in isolating individuals for whom the standardisation model has not been appropriate. As discussed below, had the process continued as initially planned, this information would have been made available through the proposed appeals process.

4.1 Summary measures

To demonstrate the effect of applying the different sets of criteria described in Section 3, a breakdown of the number and percentage of entries[15] is presented here. The total number of A level entries contained in the dataset analysed from summer 2020 is 738,418. When performing the analysis across the historical data, the total number of entries was 2,506,620 over the previous 3 years with 1,846,646 being matched to prior attainment.

4.1.1 Prior attainment-based filtering

Table 4 contains a breakdown of the number entries that remain following application of the different filtering criteria where the primary filter was based on students' prior attainment.

This is provided for the 2 versions of these analyses discussed above; those where the prior attainment z-score is evaluated solely on the basis of students with entries in the current year ('2020 only') and those where the z-score is evaluated considering students from across the current year and all 3 years of historical data used for the purposes of establishing the statistical model ('All historical data').

As reflected in the description in Section 3.1, there is a hierarchy to the filters. The primary basis for identifying outlying students in this instance was prior attainment. Shown in row a of the table is the number of entries from students who can be matched to their prior attainment with row b showing the number of entries remaining once the z-score criterion has been applied. This demonstrates that, applying the z-score threshold retains 2.1% and 2.6% of entries for the 2020 and all-year datasets, respectively.

Rows c to e demonstrate the impact of applying secondary criteria (separately and in combination) to the sub-set of entries identified as potentially outlying based on prior attainment. It is interesting to note the filtering effect of selecting only those entries for which there is a difference between the CAG and calculated grade. Of the entries identified as outlying on the basis of prior attainment 64.4% (9,981/15,505) and 68.4% (12,946/18,916) were removed as they received calculated grades that matched the CAG, for 2020 and all-year datasets, respectively. Across the whole population, the percentage of A level entries where the calculated grade matched the CAG was just under 60%. It is interesting to note that, the CAG to calculated grade match rates for this subset of students was higher than for the overall A level population. This could be interpreted in 1 of 2 ways. Either the model was reasonably robust to the handling of students that were outlying in terms of their prior

[15] While references to this point have been to students, it should be noted that the analyses are performed at the entry level.

attainment, and so with this group was not subject to additional uncertainty and potential disadvantage, or the filtering on the basis of being outlying in terms of prior attainment is not effective in identifying entries where reduced accuracy may have occurred. In reality, both of these statements are likely to be partially true.

Table 4 Number of entries resultant from applying prior attainment distribution-based filtering. (An accessible version of Table 4 is available in csv format).

				2020 only	All historical data	Number of common entries
a)		All matched entries	Number	612,841	612,841	612,841
			Percentage of all entries	82.99	82.99	-
			Percentage of entries (matched)	100.00	100.00	-
b)	Primary filter	Z-score only	Number	15,505	18,916	11,757
			Percentage of all entries	2.10	2.56	-
			Percentage of entries (matched)	2.53	3.09	-
c)	Secondary filters	With difference between CAG and calculated grade	Number	5,524	5,970	3,865
			Percentage of all entries	0.75	0.81	-
			Percentage of entries (matched)	0.90	0.97	-
d)		Within top 2 ranks	Number	7,850	10,220	5,993
			Percentage of all entries	1.06	1.38	-
			Percentage of entries (matched)	1.28	1.67	-
e)		Combined	Number	2,485	2,644	1,725
			Percentage of all entries	0.34	0.36	-
			Percentage of entries (matched)	0.41	0.43	-
f)	Optional filters	With unique CAG	Number	911	958	643
			Percentage of all entries	0.12	0.13	-
			Percentage of entries (matched)	0.15	0.16	-
g)		With lower than average generosity	Number	903	994	664
			Percentage of all entries	0.12	0.13	-
			Percentage of entries (matched)	0.15	0.16	-
h)		Combination of all criteria	Number	352	377	265
			Percentage of all entries	0.05	0.05	-
			Percentage of entries (matched)	0.06	0.06	-

The combined effects of applying the secondary filters are indicated on row e. This provides the first meaningful attempt to identify outlying entries based primarily on the prior attainment distribution within the centre for the subject. Between 0.3% and 0.4% of the cohort are identified.

Rows f and g show the impact of applying the additional, optional, filters. The first identifies only those entries where the student is the only one with that CAG. While logically necessary for a student to be considered to be outlying, this is a more refined and particularly stringent variant of the requirement for the student to be ranked in the top 2. This demonstrates a significant (63.3% and 63.8%) reduction in entries from those identified in row e. A similar reduction (63.7% and 62.4%) is achieved through the exclusion of centres with a level of generosity greater than average (row g). Applying these additional steps seeks to improve the detection of outlying students where they have been obscured in the CAG distributions of more generous centres.

Applying all of the filters simultaneously leads to the identification of 0.05% of the cohort as potentially outlying.

Before inspecting the details of these cases, it is useful to consider other features of the subsets of entries identified as the analysis is based on entries – individual students entering for individual subjects. Usually, students enter exams for multiple subjects leading to multiple simultaneous entries. This means it is possible for an individual student to occur multiple times within the filtered sets. Arguably, these instances may indicate a stronger likelihood of the student being outlying. They may also indicate a greater need for redress, particularly given the risk of multiple points of disadvantage to the student. Shown in Table 5 is a student level analysis equivalent to the entry-level analysis presented in Table 4.

In the dataset used for the 2020 analyses there were 286,225 students making up the 738,418 entries. Of these students, 229,589 were successfully matched to their prior attainment in line with the matching criteria provided in Section 3.1**Error! Reference source not found.**. As can be seen from row a of Table 5, 187,430 of those students entered for multiple subjects corresponding to 82% of students present at that point through the filtering process. Figures indicating the proportion of duplicate students are reported at each step of the filtering process.

Notably, on row e, following application of the first set of criteria that isolate a meaningful subset of potentially outlying students, 2,258 were identified from the *2020-only* analysis with only 9.2% of those students being multiply identified. This corresponds to 0.8% of the overall cohort. For the *all-year* analysis, this isolated 2,337 students with 278 or 11.9% being identified more than once.

Row h shows the results of applying all criteria with the 2020 only analysis identifying 346 students and the all-year analysis identifying 367. Within these groups only 6 and 10 students respectively were identified on multiple occasions.

These analyses show only a very small proportion of students were consistently identified as potentially outlying across subjects, however, there is still significant uncertainty regarding the correctness of their classification as outliers.

Table 5 Number of students and students with multiple entries resultant from applying prior attainment distribution-based filtering. (An accessible version of Table 5 is available in csv format)

				2020 only	All historical data
a)		All matched students	Number of students	229,589	229,589
			Percentage of all students	80.21	80.21
			Percentage of students (matched)	100.00	100.00
			Number of duplicate students	187,430	187,430
			Duplicate students (%age of subset)	81.64	81.64
b)	Primary filter	Z-score only	Number of students	10,309	11,305
			Percentage of all students	3.60	3.95
			Percentage of students (matched)	4.49	4.92
			Number of duplicate students	3,632	4,843
			Duplicate students (%age of subset)	35.23	42.84
c)	Secondary filters	With difference between CAG and calculated grade	Number of students	4,761	4,966
			Percentage of all students	1.66	1.73
			Percentage of students (matched)	2.07	2.16
			Number of duplicate students	693	871
			Duplicate students (%age of subset)	14.56	17.54
d)		Within top 2 ranks	Number of students	6,073	7,191
			Percentage of all students	2.12	2.51
			Percentage of students (matched)	2.65	3.13
			Number of duplicate students	1,445	2,259
			Duplicate students (%age of subset)	23.79	31.41
e)		Combined	Number of students	2,258	2,337
			Percentage of all students	0.79	0.82
			Percentage of students (matched)	0.98	1.02
			Number of duplicate students	208	278
			Duplicate students (%age of subset)	9.21	11.90
f)	Optional filters	With unique CAG	Number of students	865	895
			Percentage of all students	0.30	0.31
			Percentage of students (matched)	0.38	0.39
			Number of duplicate students	45	57
			Duplicate students (%age of subset)	5.20	6.37
g)		With lower than average generosity	Number of students	869	948
			Percentage of all students	0.30	0.33
			Percentage of students (matched)	0.38	0.41
			Number of duplicate students	33	45.00
			Duplicate students (%age of subset)	3.80	4.75
h)		Combination of all criteria	Number of students	346	367
			Percentage of all students	0.12	0.13
			Percentage of students (matched)	0.15	0.16
			Number of duplicate students	6	10
			Duplicate students (%age of subset)	1.73	2.72

4.1.2 CAG based filtering

Table 6 contains a breakdown of the number entries that remain following application of the different filtering criteria where the primary filter was based on students' prior attainment. In this instance, the preliminary filter is based on identifying entries whose CAG z-score is greater than 2.0. Subsequent filters were then separately applied to the sub-set of entries identified through the initial filter, before applying all criteria simultaneously.

Table 6 Number of entries resultant from applying CAG distribution-based filtering. (An accessible version of Table 6 is available in csv format)

				2020 only
a)	Primary filter	Z-score only	Number	4,995
			Percentage of all entries	0.68
b)	Secondary filters	With difference between CAG and calculated grade	Number	2,763
			Percentage of all entries	0.37
c)		Within top 2 ranks	Number	4,292
			Percentage of all entries	0.58
d)		Combined	Number	2,272
			Percentage of all entries	0.31
e)	Optional filters	With unique CAG	Number	1,245
			Percentage of all entries	0.17
f)		With lower than average generosity	Number	981
			Percentage of all entries	0.13
g)		All criteria	Number	509
			Percentage of all entries	0.07

As the CAGs are not comparable with the awarded grades in previous years[16] it is not appropriate to consider distributions of CAGs along with previous actual grades in the way that was possible with the measures of prior attainment. The analysis in this section, therefore, only includes data from summer 2020.

The first point to note is the proportion of entries retained through the CAG z-score filter compared to the prior attainment-based filter with only 0.7% of entries occurring above the *Z* = 2.0 threshold. This is lower than the theoretical proportion (2.3%) and may be an artefact of assuming a normal distribution for what is a highly discrete variable.

[16] As reported in Section 9.1 of the interim report, the outcomes based on CAGs were notably higher than in previous years. For example, aggregated across all A levels, the outcomes at grade A based on CAGs was 12.5 percentage points higher than the results in 2019.

Also of note is the proportion of these most extreme outlying students for who the calculated grade was the same as the CAG. This is shown in row b of the table with 2,232 entries (or 44.7% of the outliers based on z-score) being removed due to this match. This is a lower rate of matching compared to the prior attainment-based filter and compared with the average across the whole cohort. This may suggest that this is a more effective approach to identifying students for whom standardisation has been less effective. Alternatively, these CAGs might be outlying due to particularly generous centre judgements. This would make the adjustments appropriate. This again highlights the limitations of purely statistical approaches to identifying outlying students and the risks of mis- or over-interpretation without additional evidence.

Row c shows that relatively few entries are removed by filtering out entries where the student is not ranked in the top 2 for the subject. Only 14% (703/4,995) of entries were removed by this step. However, given the interaction between the CAG based z-score filter and the rank order of students, this insensitivity is unsurprising.

In conjunction, this leads to 0.3% of entries being identified as potentially outlying (row d). Having applied the 2 additional filters in row g this identifies just 509 entries (0.07%).

Table 7 shows the analysis of duplicate students within the entries dataset. This seeks to identify instances where there is risk of multiple disadvantages to students.

Application of the primary CAG filter on row a shows a notably lower rate of duplicate students (8.0%) compared to the equivalent point in the prior attainment-based filtering (row b) in Table 5; 35.2% and 42.8%). This is likely to be for 2 reasons. First, the CAG-based approach leads to a notably lower proportion of entries being identified through this first stage compared with the prior attainment-based approaches (0.7% compared 2.5% and 3.1%). This means that there is a lower likelihood of a student appearing in the filtered dataset purely due to chance. Second, the CAG is a subject specific measure meaning that it may, and in many cases will, vary across a student’s subject entries. While the distribution against which a student’s prior attainment measure is compared varies across subjects, the measure itself remains unchanged.

Table 7 Number of students and students with multiple entries resultant from applying CAG distribution-based filtering. (An accessible version of Table 7 is available in csv format)

				2020 only
a)	Primary filter	Z-score only	Number of students	4,595
			Percentage of all students	1.61
			Number of duplicate students	368
			Duplicate students (%age of subset)	8.01
b)	Secondary filters	With difference between CAG and calculated grade	Number of students	2,623
			Percentage of all students	0.92
			Number of duplicate students	131
			Duplicate students (%age of subset)	4.99
c)		Within top 2 ranks	Number of students	3,994
			Percentage of all students	1.40
			Number of duplicate students	274
			Duplicate students (%age of subset)	6.86
d)		Combined	Number of students	2,168
			Percentage of all students	0.76
			Number of duplicate students	97
			Duplicate students (%age of subset)	4.47
e)	Optional filters	With unique CAG	Number of students	1,198
			Percentage of all students	0.42
			Number of duplicate students	44
			Duplicate students (%age of subset)	3.67
f)		With lower than average generosity	Number of students	959
			Percentage of all students	0.34
			Number of duplicate students	22
			Duplicate students (%age of subset)	2.29
g)		Combination of all criteria	Number of students	496
			Percentage of all students	0.17
			Number of duplicate students	13
			Duplicate students (%age of subset)	2.62

4.1.3 Inter-analysis comparisons

To attempt to further validate the identification of entries as outliers it is informative to compare across the filtered datasets. The occurrence of an entry across multiple analyses may suggest that they are indeed outlying. Provided in Table 8 is a breakdown of the number of entries that are common across the sub-sets produced at the different stages of filtering. To provide context for these figures, percentages are quoted relative to the number of entries identified at the equivalent stage of the process in the prior attainment all-years analysis.

Table 8 Cross-analysis comparison of identified entries. (An accessible version of Table 8 is available in csv format)

			Prior 2020 v Prior all-years		Prior 2020 v CAG 2020		Prior all-years v CAG 2020	
			Number	as %	Number	as %	Number	as %
a)	Primary filter	Z-score only	11,757	62.2	1,555	8.2	1,396	7.4
b)	Secondary filters	With difference between CAG and calculated grade	3,865	64.7	817	13.7	679	11.4
c)		Within top 2 ranks	5,993	58.6	1,390	13.6	1,237	12.1
d)		Combined	1,725	65.2	701	26.5	574	21.7
e)	Optional filters	With unique CAG	643	67.1	432	45.1	357	37.3
f)		With lower than average generosity	664	66.8	324	32.6	285	28.7
g)		Combination of all criteria	265	70.3	196	52.0	174	46.2

The first point to note is the reasonably high level of commonality between the sub-sets produced by the first step of the prior attainment-based approaches with 62% of entries being common. Although this level of commonality may be considered relatively modest given the similarity of the approaches. This reasonably high level of commonality is retained through the different stages of filtering. Again, this is unsurprising and not necessarily informative given the high levels of similarity between the methods with the only difference being the population of students making up the analysis in the primary filter.

The low levels of commonality between the CAG-based filtering and the prior attainment-based approaches are notable, however. Match rates with the prior attainment-based approaches are only 8.2% and 7.4%. This is likely to be partly due to the limitations of using prior attainment as a basis for the filtering. It reflects the fact that having outlying prior attainment does not necessarily indicate the student will be outlying in terms of *current* attainment in a specific subject.

Given the identical nature of the criteria beyond the initial primary filter, the relative increase in commonality through rows d and g is inevitable and these sub-sets of entries become similar, not due to the accuracy of the primary basis on which entries are identified as outliers but based on the additional criteria included to verify the datasets.

4.2 Case studies

All of the above analyses have been focused on quantitative indicators relevant to the assumptions made by the statistical standardisation model. It is not possible, however, to validate the results of these filtering processes without inspecting the outcomes and, as far as is possible, evaluating the plausibility of the identified cases being outlying.

Therefore, a small number of case studies are selected to support the discussion.

Shown in Table 9 and Table 10 are the summary measures which characterise particular entries and information about the distributions of the centre's cohort for the subject. These are provided for cases identified through the prior attainment (*2020-only*) and CAG based criteria, respectively. Included in these tables are only those entries that meet the criteria for the primary and secondary filters (i.e the relevant z-score threshold, a difference between CAG and calculated grade, and ranked within the top 2 for the subject within the centre). A subset of those selected also meet the optional criteria as indicated by †.

Notes summarising these cases are provided along with the data entries in the tables and, therefore, is not repeated here. What is clear, however, is that even when the more stringent optional filtering is applied, there are many cases where the is notable uncertainty regarding whether the students identified have been disadvantaged through the standardisation process and, even where this appears to be the case, what the awarded grade should be.

This section has considered the application of the criteria described in section 3. This has shown that the different combinations of criteria give rise to different sub-sets of students being identified, however, the confidence in those criteria not producing significant numbers of false-positives and false-negatives is very low. Even when visually inspecting individual instances, in many cases, it is not possible to determine whether or not the assumptions of the model have been violated leading to an unreliable result. Also, the issue of how to remedy any instances where outlying students have been successfully identified is not trivial and is the subject of the section that follows.

Table 9 Case studies arising from the prior attainment-based criteria. (An accessible version of Table 9 is available in csv format)

Case	Individual measures					Centre-level measures																Notes
	z-score	CAG	Calc grade	Unique on CAG	Rank	Entries	CAG distribution							Calculated grade distribution							Mean generosity	
							A*	A	B	C	D	E	U	A*	A	B	C	D	E	U		
1	2.85	A*	C	Y	1	37	1	6	11	9	5	4	1	0	0	0	12	18	5	2	1.19	A large (3 grade) difference between CAG and calculated grade, but high-level of generosity in the CAGs compared with the statistical prediction (1.19 grades per entry). The calculated grade being lower than the CAG appears appropriate, but the appropriate magnitude of the difference unclear.
2	2.70	A*	B	N	2	19	5	2	5	4	2	1	0	0	1	7	7	4	0	0	0.79	The apparent generosity in the CAGs relative to the calculated grades is higher than average. The student is ranked 2nd within the centre. While it appears reasonable that the calculated grade is lower than the CAG, it is uncertain whether a difference of 1 or 2 grades is appropriate.
3	3.58	A*	A	N	1	44	5	15	17	7	0	0	0	0	3	15	18	6	0	2	1.20	There is a difference of a single grade the CAG and calculated grade. Given the high level of apparent generosity of the centre (1.20 grades), the calculated grade being 1 grade lower than the CAG (along with the other entries with CAG = A*) appears appropriate and consistent with the available evidence.
4	3.30	A*	A	N	1	25	2	0	10	7	3	3	0	0	2	8	9	4	2	0	0.12	The CAGs appear only very slightly generous compared to the statistical prediction. The downward adjustment of this outlying student, therefore, appears to represent potential disadvantage.
5	3.29	A*	A	Y	1	22	1	4	6	10	1	0	0	0	1	3	7	7	3	1	1.23	There is a difference of a single grade the CAG and calculated grade and the student has a unique CAG. Given the high level of apparent generosity of the centre (1.23 grades), the calculated grade being 1 grade lower than the CAG this adjustment appears consistent with the available evidence.
6†	3.27	A*	A	Y	1	20	1	0	6	10	2	1	0	0	1	6	9	4	0	0	0.05	The CAGs appear only very slightly generous compared to the statistical prediction. The downward adjustment of this outlying student, who is also an outlier based on their CAG, therefore, appears to represent potential disadvantage.
7†	3.26	A*	A	Y	1	21	1	1	7	8	4	0	0	0	2	6	6	5	2	0	0.33	The level of apparent generosity in the CAGs is below average. It is unclear from this evidence whether the adjustment of the top most student is appropriate.

Table 10 Case studies arising from the CAG based criteria. (An accessible version of Table 10 is available in csv format)

Case	Individual measures					Centre-level measures																Notes
	z-score	CAG	Calc grade	Unique on CAG	Rank	Entries	CAG distribution							Calculated grade distribution							Mean generosity	
							A*	A	B	C	D	E	U	A*	A	B	C	D	E	U		
1†	2.81	A*	B	Y	1	12	1	0	0	2	7	2	0	0	0	2	2	4	3	1	0.25	There is a larger than typical difference between the CAG and calculated grade (2 grades), however, the level of generosity that appears to be present in the CAGs is modest. On this basis, the difference appears anomalous and the student may have been potentially disadvantaged.
2	2.21	A*	B	N	2	33	2	6	11	13	1	0	0	0	0	2	11	14	5	1	1.61	The level of generosity in the CAGs compared to the calculated grades appears to be significant (1.61 grades per student). On this basis, the adjustment applied appears necessary, however, it is unclear whether the level of adjustment is appropriate.
3	2.59	A*	A	N	1	86	2	9	19	40	12	4	0	0	2	12	22	29	15	6	0.98	The difference between the CAGs and the calculated grades is higher than average. This feature, combined with the nature of the CAG distribution in comparison with that of the calculated grades, suggests that the award of grades was likely appropriate.
4†	3.11	A*	A	Y	1	23	1	0	1	12	8	0	1	0	2	4	7	9	1	0	-0.17	The overall accuracy of the CAG distribution compared with the calculated grades would suggest the distribution is broadly legitimate. In addition, a student receiving a calculated grade A having received a CAG of B suggests that the same grade being awarded to the student with a CAG of A* is probably inappropriate.
5†	2.58	A*	A	Y	1	45	1	4	15	16	8	1	0	0	2	12	16	11	3	1	0.44	The level of generosity in the CAGs is average. The award of a calculated grade A rather than the CAG of A* appears plausible and not anomalous, but there is some uncertainty as to its appropriateness.
6	2.55	A*	A	N	1	108	2	14	34	40	17	1	0	0	6	38	46	15	2	1	0.19	Both students ranked 1 and 2 had the same z-score as they shared a CAG. It is unclear whether just the lower ranked student, both students or neither students should receive the CAG or calculated grade.
7	2.55	A*	A	N	2																	

5. Remedy and appeals

Irrespective of the approach taken to identify outlying students, commonly in statistical analyses, the course of action once they have been detected is the same; they are removed from the analysis and – with the exception of noting their existence when reporting results – they are typically ignored. This course of action is usually entirely fitting and appropriate when the purposes of the analyses are to identify or explore some underlying relationship as part of a research study. In the context of generating results for individual students, however, this approach is clearly not appropriate, since every student in the dataset must be allocated a grade. Ignoring or removing a student from the process because they may be statistically atypical would clearly be both unfair and unacceptable. It is necessary to consider, therefore, how such issues might be resolved where outlying students have been identified.

As has been demonstrated above, identifying students as outlying for who the standardisation model has not reliably functioned is complex and cannot be satisfactorily achieved through the use of objective *a priori* measures available at the time of standardisation. Consequentially and most unfortunately, these issues could not be resolved in advance of grades being issued to students. It was necessary to put in place a suitable appeal process in which new information could be brought to bear. This process would be supported by but not wholly reliant on, the same quantitative information used in the standardisation model.

The appeal arrangements that were put in place are detailed in our regulations. In summary, among other issues, these arrangements were principally seeking to facilitate deeper evidence-based discussions of individual students or groups of students where there was quantitative evidence suggesting the statistical model may have been unreliable.

Where the appeals process identified students for who the standardisation model was unreliable, it would have been necessary to determine an appropriate remedy - to identify an alternative ‘correct’ or more correct grade. In the vast majority of cases where there was a difference between the CAG and calculated grade, that difference was a single grade. This was the case for 92.9% (5,367 of 5,780) of entries making it through the preliminary filtering based on prior attainment for 2020 only, 93.9% (5,879 of 6,259) when using prior attainment from recent years and 87.0% (2,440 of 2,806) of entries making it though the preliminary filter based on CAGs where the CAG and calculated grade were different. In these instances, where the appeals process has identified an unreliable outcome, identifying the grade to be awarded is obvious as, once the calculated grade has been dismissed, the CAG is the only logical result.

There will be cases, however, where the difference between the CAG and calculated grade is more than a single grade. In these instances, the most appropriate result is less-obvious. For the purposes of discussion, such an example is illustrated in below.

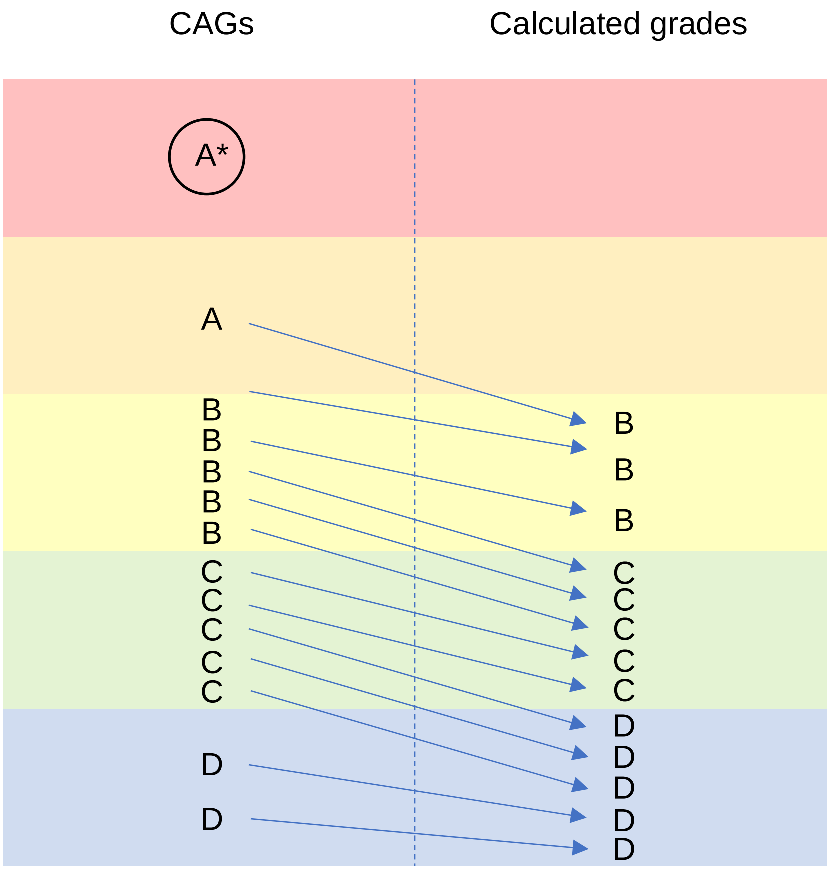

Figure 2 Illustrative example of uncertainty in resolving the most appropriate grade for an outlying student

Shown in the figure are the students' grades based on CAGs and the calculated grades resulting from standardisation. Here, it is assumed that the highest-ranking student (circled) has been identified as being outlying through the appeals process and supporting analyses. It is assumed that the results for the remainder of the cohort were deemed appropriate despite the calculated grades being lower than the CAGs. The outlying student was awarded a calculated grade B.

In this scenario it is impossible to know whether the overall generosity apparent in the CAGs was also present in the judgement regarding the likely performance of the outlying student. It is difficult to determine whether the outlying student should be awarded a grade A (taking into account the overall generosity) or whether the student was so able that they should be awarded a grade A*, irrespective of the overall generosity of judgement.

There are a multitude of similar but different scenarios. Rather than considering these circumstances individually, there is also a question of principle that potentially resolves the issue. By definition, when a student has been deemed to be outlying it has been acknowledged that the standardisation model cannot reliably award their grade. In this situation it would be illogical to take into account any information provided by the model in relation to this student in this subject. At this point, the decision making should default to the best available evidence. Irrespective of any uncertainty in the judgements leading to the CAGs (in the example given above, due to the apparently leniency of judgements relating to other students), but in the presence of even greater uncertainty about the statistical evidence for this student, it would seem appropriate on principle to award the CAG, where evidence has become available through the appeals process.

5.1 Alternative appeal scenarios

For the purposes of simplicity and reflecting the focus of public interest in issues relating to outlying students, the majority of the discussion provided here has been centred on outlying students at the top of the ability range. Similar challenges also exist, however, at the bottom of the grade distribution. In the simplest terms, all the discussions relating to particularly able outlying students can be mirrored to the consideration of the least able students. There is, however, a marked difference between the 2 scenarios insofar as those students would be advantaged through the process and, therefore, the negative consequences for the individual are less apparent. There are, however, further assumptions at the bottom of the grade distribution that may negatively impact on outcomes for students. In contrast to the consideration of outlying students discussed above where there may be an atypical *presence* of students, this would occur where the distribution of abilities within a centre was atypical due to an *absence* of students. It is worth considering whether the principle of accepting CAGs where the statistical model has been deemed to be inappropriate for the student is appropriate in such instances. To support this discussion, an example of this is illustrated Figure 3 in below.

Here it is assumed that it has been confirmed through the appeals process that the assumptions of the model have been violated for a sub-set of the students. In this example, the 2 circled students have been deemed to have been unfairly disadvantaged by the standardisation process whereas the award of grades to the remainder of the cohort is considered to be appropriate.

In such a situation, the principle suggested above of awarding the CAGs to the outlying students in instances where the statistical model has proved inappropriate, would overwrite some potentially stronger judgemental evidence from the centre. As can be seen from the diagram, the CAGs for the 2 outlying students are grade C with the calculated results being grade E and ungraded. Reinstating the CAGs to these students and awarding them a grade C, however, would disrupt the rank order submitted by the centre as this would lead to them being awarded a higher grade than students legitimately being awarded a calculated grade D.

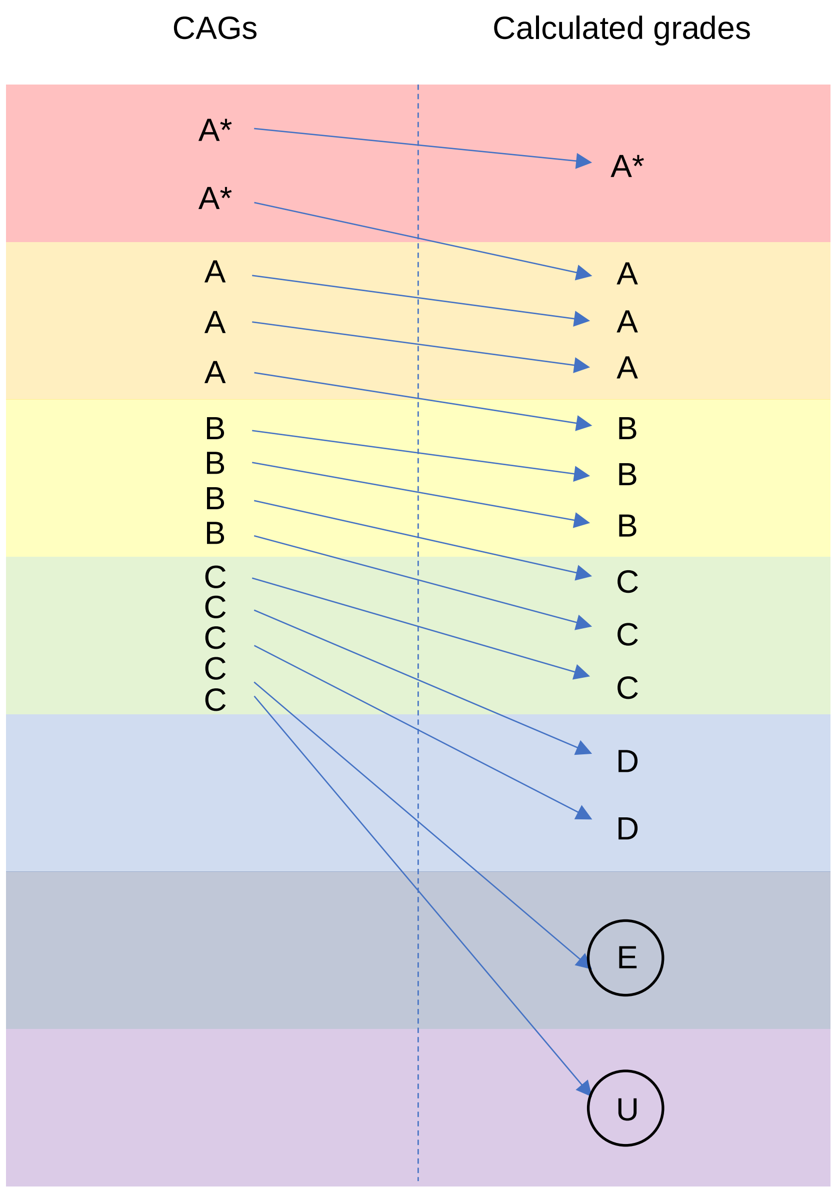

Figure 3 Illustrative example of a centre with an atypical distirbution of students at the bottom of the distribution

To accommodate such a situation it is, therefore, desirable to modify the principle suggested above of awarding the CAGs where it is judged that the statistical model is inappropriate for certain individual students. To protect the rank order provided by centres and to treat students across the cohort as fairly as possible it is suggested that the most appropriate approach would be to award whichever is lower between the CAG and the calculated grade of the next student above in the rank order for whom the statistical model has been deemed to be sufficiently reliable. In the example provided in Figure 3, this would result in both students highlighted being awarded a grade D.

Based on the uncertainty with which it is possible to identify such cases based on *a priori* information and the practical limitations of delivery, reliable identification could only be performed through a post-results appeals process.

6. Summary

To enable the award of grades in general qualifications in summer 2020, centres submitted the grades they estimated their students would have achieved had exams taken place and the rank order in which they expected students would have performed. To mitigate the risk that centres might apply differential standards when coming to their judgements, which would be unfair to students, we developed a statistical model to standardise the grades awarded. The model predicted the centre's grade distribution in a subject based on the historical performance of the centre in that subject and the differences between the prior attainment of this year's student cohort compared to previous years.

As with any statistical model, the standardisation model made assumptions about groups of students and inevitably these assumptions did not hold for every student. In the lead up to the issuing of results and immediately following A level results day, these students were often referred to as 'outliers' in the media. There was significant concern that the model would be unable to accommodate outlying students and that this might mean they would be awarded unreliable grades.

The definition of outliers in the context of the standardisation model does not follow the conventional statistical definition. Typically, outliers are defined and detected based on an observed data point appearing to be anomalous or, at least, probabilistically unlikely. In relation to standardisation, this would correspond to a student receiving a calculated grade that was significantly different to their CAG. However, concerns regarding outlying students often related to students who were atypical within their centre. Unreliable results for these students would not necessarily be observed as a large difference between the CAG and the calculated grade. Given the impact on students' life chances, students who could be defined as outlying based on their input characteristics, for whom there was a small difference between CAG and calculated grade were of as just as much concern as the relatively small number where the difference was large.

To identify outlying students, criteria were designed to be sensitive to either the prior attainment or the CAG of each student in relation to the other students taking the subject in that centre. This assumed that outlying students would either be atypical based on their prior attainment (meaning their presence may be insufficiently compensated for in the model) or their CAG (meaning they may be assumed to fit with the distribution of grades for the rest of the cohort when they should not). Additional requirements were set as part of those sets of criteria to reduce the rate of false positives.

Analyses show that the success of these quantitative criteria to identify outlying students with unreliable grades was mixed. Inspection of the flagged cases threw up examples where it seemed likely that the statistical model had not resulted in a defensible grade. Equally, however, there were examples where the operation of the model appeared appropriate or where it was not possible to determine whether or not the potential outlying student had indeed received a grade that was less reliable than others.

Despite being less statistically sound, evidence suggests that the CAG-based criteria may have been better at identifying outlying students than criteria based on prior attainment. One indication of this is the rate at which the prior attainment-based criteria identified students whose CAG was identical to their calculated grade. This was the case for 64% of entries with a z-score greater than 2 when using data from summer 2020 only and was 68% when the analysis included data from previous years. This match rate is higher than that which was observed across the whole A level cohort (59%). This suggests that students with outlying measures of prior attainment were more likely to receive a grade aligned to their teacher’s expectation rather than less likely. In contrast, of the entries identified through the CAG-based criterion, 45% of entries had CAGs which were equal to the calculated grades.

The direct nature of the CAG-based approach was also attractive. Standardisation models that relied on the characteristics of individual students such as prior attainment were dismissed through the model design process due to the risks of unreliability of these measures at an individual level. This decision appears to be further supported by this evidence.

Despite the seemingly better performance when the detection of outlying students is based on the distribution of CAGs, there were still significant limitations to the results produced. As described in Section 4.2 and shown in the presented case studies, while some instances provide seemingly clear evidence that the results for identified students may have been anomalous, there are many instances where this is not the case and the correct course of action is either uncertain or the functioning of the standardisation model appears to have been effective.

Ultimately, based on the criteria explored, we decided it would be inappropriate to build into the standardisation process a mechanism to attempt to identify outlying students and to compensate for their presence. This was because of the lack of confidence in the effectiveness of the identification criteria and the largely arbitrary nature of those criteria. Were such an approach to be taken, these 2 factors would combine to risk of additional unfairness. This is because a sub-set of students would be incorrectly identified as outlying when in fact the model was appropriate and another sub-set of outlying students who had received unreliable grades would fail to be identified.

In addition, there are many instances where the appropriate redress for the outlying student is not clear.

We, therefore, decided that any unreliable grades for outlying students would be best addressed through an efficient post-results appeal process. Taking this approach would facilitate an in-depth consideration of technical information and richer context-specific evidence which could be provided by the centre. This would allow corrected grades to be calculated, informed by a solid evidence base.

In fact, the appeals process was not used because, following the issue of standardised A level results, it became apparent that the grades issued did not command public confidence. A great deal of distress was experienced by students, their families, teachers and the wider public and for this we are very sorry. In light of this anguish, we decided to award grades to students that were either the CAGs or the standardised calculated grade, whichever grade was higher.

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344

public.enquiries@ofqual.gov.uk

www.gov.uk/ofqual

December 2020 **Ofqual/20/6732**