Awarding grades at GCSE and A level – a response to Ofqual's consultation

Summary

Ofqual and the Awarding Bodies face an extremely difficult task. In a year where examinations are not being sat, achieving complete fairness in grading will be impossible. Nonetheless, society owes it to the affected young people not to disadvantage them further in a situation not of their making.

Ofqual and the Awarding Bodies should recognise that the harm of awarding an unfairly low grade this year is greater than the harm of awarding an unfairly high grade and so adopt a 'precautionary principle', balancing the need to maintain standards over time against the need to avoid unfair prejudice to the individual. This may require some easing of the 'comparable outcomes' distribution and possibly some adjustment to the policy framework within which Ofqual operates.

Ofqual has recognised the problems of teacher-assessed grades but may over-state them because incentives on schools and colleges have been removed by the removal of performance tables. Ofqual may, however, over-estimate the extent to which statistical methods using historic data can eliminate variation in standards applied by centres. There is no statistical method for distinguishing between genuine large changes from one year to the next and misapplication of the standard.

A rigid application of a statistical model to all centres could itself introduce substantial unfairness to individual exam candidates. Ofqual should take more account of school assessed grades and accept these unless there is reasonable evidence to suggest that they are not accurate. Ofqual and the Awarding Bodies can take steps in the coming month to improve the likelihood of centre assessed grades being accurate – including guidance and training on how centres should use data to support judgements, and clear accountability for centres, whose evidence should be open to scrutiny.

Ofqual should consider adopting a process in which decisions about the approach to moderation in a subject are dependent on the actual distribution of centre assessment grades Awarding Bodies receive. Where the national distribution of centre assessed grades in a subject is not unreasonably far from an expected distribution, Ofqual should accept it, adjusting only for centres where there is clear evidence of inaccuracy. Where the proposed distribution is too far from expected to be acceptable, Ofqual should approach standardisation by looking for confirmatory evidence rather than immediately fitting centres to a single pre-determined distribution and should use the widest possible range of data and evidence in doing so.

In due course, Ofqual should make public its models for standardisation and be prepared to explain and justify what their impact would have been had they been applied to prior year results. Standardisation models may need to be different subject-by-subject to achieve the fairest result, and Ofqual should be unconcerned about the complexity this will introduce. Ofqual should consider whether looking for centre-level assurance by examining data across a range of subjects could provide a basis for identifying legitimate outliers at subject level.

Ofqual should also re-consider whether there is a case for examining further evidence from centres where there are significant changes in the results which could not be predicted by historic data. Statistical methods can identify apparently anomalous results but cannot tell us whether these results are genuine outliers arising from correct application of the process or examples of inaccurate assessment. Where a centre believes that performance of students in a subject or range of subjects has changed markedly, it should be able to put forward for assessment evidence to support this.

There is no perfect methodology. However, we think that some easing of the standard and placing of weight on centre assessed grades could make substantial injustice to young people less likely.

Introduction

In a year where GCSE and A level examinations will not be taken, the task of awarding examination grades is immensely difficult. It will not be possible to achieve as fair a result for young people as could be achieved were examinations sat – if it were, the annual examinations process would not be needed. Undesirable though this situation is, it is the product of a health emergency unprecedented since the introduction of mass public examinations and not the fault of Ofqual or of the Awarding Bodies, who have an extremely difficult job to do in this context.

We should acknowledge that complete fairness is impossible and focus effort on protecting this year's candidates

The fact that complete fairness cannot be achieved should be acknowledged openly. It is particularly important this year that qualification 'end users' (including universities, post-16 educational institutions and employers) are encouraged not to place undue weight upon small differences in grades achieved. This is always dangerous in any case, but much more so in a year when final grades will not be supported by wholly consistent and independent assessments.

In this context, society and the exam system have a particular responsibility to this year's candidates to avoid harming their life chances. This year's students – uniquely – have been prevented from sitting exams which they have spent some years working towards. Furthermore, the need to suspend schools arose not from substantial threat to the majority of the young people affected by that suspension but from a need to protect older adults.

This situation creates an asymmetry to the risks in the awarding process which does not exist in a normal year. The risk of harm arising from the award of a grade which is lower than that which a student would otherwise have achieved is much greater than the risk of harm arising from awarding a grade which is higher than would otherwise have been achieved. A student unfairly missing grades may miss out on a place at University or the opportunity to proceed to sixth form or to level 3 courses, which may have life-changing consequences. GCSE students not achieving grade 4 in English or maths will be required to re-sit. If a student unfairly achieves a higher grade, that will have at most marginal impact on other students in their year or other years.

The need to maintain standards over time must be balanced with the need to avoid prejudice to the individual exam candidate

In a situation where the usual level of fairness is unachievable then there will be 'winners' (those who get higher than deserved grades) and 'losers' (those who get lower than deserved grades). If the national approach to standard-setting is maintained to give precisely the same proportion of each grade as would be expected under 'comparable outcomes', then arithmetically, there will be precisely the same number of winners and losers (though we cannot know the number of either).

If it is accepted that the unfairness is greater (and the consequences potentially substantially more severe) of having 'losers' in this system than 'winners', then this is not the optimal outcome. A better outcome would be achieved through some easing of 'comparable outcomes' in order to reduce the number of losers at the cost of having slightly more winners.

That is not to say that great efforts should not be made to make the process as fair as possible. A dramatic change in exam standards in one year could not be tolerated and would undermine the value of grades achieved by the whole cohort. However, if changes are made to the standard at the margins in order to protect individuals because certainty of grade is unachievable, then this is defensible and will be widely understood in the unique circumstances of this year.

Ofqual and the DfE should agree an amendment to Ofqual's remit in relation to standards over time this year, altering the regulatory framework if necessary. This should require Ofqual to balance the twin requirements of protecting standards over time with the need to avoid significant unfair prejudice to the individual.

The proposed autumn exam series is welcome but does not remove the negative impact on individual candidates of unfairly low grades in the summer

The proposal to introduce an autumn sitting for students unhappy with the grades awarded is welcome and sensible. But it should not be imagined that this solves the problem for students awarded unfairly low grades.

In the first place the timing of the series and results means that students who missed the grades they need for University or sixth form will not receive a new grade until it is too late for them to progress seamlessly to their preferred destination. Students will face a choice between taking a year away from education or proceeding to a destination less attractive to them.

Secondly, for most students the current situation will hugely damage their ability to prepare for an exam. Students are away from school and studying in a new way remotely. Most schools will this term move to preparing year 11 students for post-16 study and year 13 students completing A level for HE. Realistically, current exam candidates will not be able to prepare well in several subjects for an exam series in the autumn.

Finally, preparing for an exam series in the autumn would not be cost free for the individual student. For example, a GCSE candidate who had moved onto A level in September would only be able to prepare for an autumn GCSE sitting at some cost to their A level study.

It is therefore crucial that in this summer awarding process, Awarding Bodies and Ofqual do all they can to prevent disadvantage to candidates.

The aims of standardisation should be to provide candidates with the grades they would have achieved and particularly to avoid unfairly low grades

With that in mind, we do not think that the aims of standardisation proposed in the consultation capture this need fully. Of the aims proposed on p27 of the document, aim (i) (i.e. provide candidates with the grades they would have achieved) is overwhelmingly the most important.

We do not consider that large parts of proposed aims (ii) and (iii) should be aims at all. It is not important that a common standardisation process should be used between subjects – the standardisation process used in each specification should be the one most likely to give the 'right' result and if this process is different between subjects, the desire for uniformity should not outweigh the aim of getting the best possible result.

Likewise, it is not important to use a method that is easily explained. The methods used must be made available transparently and capable of examination and challenge. They must be explained fully and should stand up to thorough scrutiny by people with the expertise and technical competence to make judgements about it. But ease of explanation to the public is of much less importance than getting the right result for the students.

Aim (iv) will inevitably be achieved if aim (i) is achieved and is highly unlikely to be achieved if aim (i) is not achieved. The process should therefore focus on aim (i). In relation to aim (v), deliverability and timeliness are of course important, but Awarding Bodies and Ofqual have some months in which to devise and implement a functional statistical process, so this should not be a major constraint.

Critically, however, the aims as set out in the consultation document do not recognise sufficiently the fundamental asymmetry described above. Where the context means that some grades will not be fair, it is substantially worse to under- rather than to over-grade a candidate. So while aim (i) as proposed is crucial, we know that it will not be possible to devise a process which perfectly awards every candidate the grade they would have secured through examination. Therefore, standardisation must also have as a specific aim the need to avoid unfair prejudice to the individual, which should be balanced with aim (i).

Ofqual is highly aware of the risks of accepting teacher/school grading but may over-emphasise these

We agree that there are risks simply to accepting teacher grades. The finding that teachers are more likely to over- than to under-grade students relative to external exams is established. It is in any case certain that in a system where thousands of people are separately making judgements, those people will not all apply precisely the same standard, however diligently they try to do so.

However, we are concerned that Ofqual are over-emphasising these risks while under-emphasising the risks of the statistical moderation process. In the past, the use of internal assessment was distorted by the fact that schools were judged using their own assessment judgements. This will not be the case this year: since no performance tables will be produced and grades will not be used for inspection schools cannot benefit from exaggerating the attainment of students. The incentives on schools which might have exacerbated concern about teacher assessment have been removed.

Furthermore, in our experience, heads, leaders and teachers are taking extremely seriously the need to get grading right and are putting in place careful quality assurance processes designed to secure accurate and evidence-based results. Schools are themselves taking careful account of past performance and using past achievement data (including raw data, prior attainment and 'transition matrices') as part of the process of predicting grades.

Where schools do put forward a different pattern of results from that which would be statistically predicted, this is likely to be fully understood by the school. Ofqual could require schools to examine certain statistical evidence as part of their internal processes in order to guarantee this. The schools' gradings will then be judgements based on evidence as to the performance of the current cohort compared to preceding ones.

We acknowledge that no matter how well this is done, it cannot guarantee perfect national consistency. However, we think that teacher grading is a more trustworthy process than the consultation document appears to suggest, while the document over-states the fairness and consistency that is likely to be achievable through statistically-based standardisation.

Ofqual may under-state the risks of statistically-based standardisation

We acknowledge the attractions of using schools' rank order as a fixed point in the system: we agree that this is likely to be sound. The finding that teachers find it easier to order than to assess against an external standard is well established and a specific example of a wider phenomenon well-established in the psychological literature.

However, we have yet to see evidence that a statistical model can accurately predict the future distribution of grades in a centre based on past performance of other students in that centre and the prior attainment of the current year's candidates. This flies in the face of our own experience: there can be substantial changes from one year to the next in cohort performance in many subjects in many schools, for reasons not explicable by historic statistics.

In our experience, this is not abnormal in schools serving areas of deprivation. Sharp falls are possible in these schools, as are rapid improvements, sometimes associated with changes of management. Such changes cannot be predicted from historic data. There is therefore a risk that a disproportionate number of those who are awarded unfairly low grades are students from deprived backgrounds, who are more likely to attend less stable schools.

Ofqual states that a downside of assuming that teacher grades are correct in the absence of statistical evidence to the contrary is that 'it is likely that differences in the standards applied by different centres would persist'. This may be true, but the document supplies no evidence that an approach to statistical standardisation between centres based on historic data could eliminate such differences in standards. Indeed, we are not aware of any such evidence and in the published data there appears to be reason to believe that rigid standardisation between centres based on historic data would introduce substantial unfairness (and the extent of this unfairness could in fact be quantified).

Ofqual's published analysis helpfully quantifies the extent of variation in cohort performance between years and some examples are reproduced as an Annex to this paper. For example, even in large entry subjects like English Language GCSE, in only about a third of schools did the proportion of students achieving grade 4+ in 2019 lie within +/-2.5 percentage points of the 2018 proportion. In a few outlying centres, the change was more than 20 percentage points – and this remains true looking only at centres of at least 100 candidates (i.e. at least 20 more or 20 fewer candidates achieving a grade 4+ than would have been predicted using the prior year's proportion). At grade 7+, the proportion within 2.5 percentage points was unsurprisingly higher, but still well below 50%.

In Art and Design GCSE and some other subjects, the variation from one year to the next is much greater. Fewer than a quarter of centres saw change in the proportion achieving 4+ from one year to the next of less than 2.5 percentage points, and some centres saw change of more than 40 percentage points. Similar levels of year on year variation are visible even in large entry A levels, including mathematics.

It may be that measured changes in prior attainment explain some of these year-on-year variations (though it is also possible that they do not). The extent to which such changes do so will presumably vary between subject and stage. It seems less likely that KS2 results (in just two subjects from five years earlier) will explain most variation at GCSE than that GCSE results in a wide range of subjects two years previously will explain a substantial part of the variation at A level. It seems less likely that KS2 results will explain variations in Art and Design results than that they will do so in mathematics.

The point of this discussion is that a rigid approach to applying a pre-determined statistical distribution to the grades achieved in a subject and in a centre will certainly create unfairness for students because the evidence shows that there is substantial variation from year to year in proportions achieving the different grades. Indeed, if the approach adopted were to reduce sharply the variation from one year to the next and so create a markedly different distribution, it would be possible to quantify quite accurately the number of students who were 'winners' and 'losers' as a result of being awarded the wrong grade (though we could not know who those students were).

<u>Ofqual's proposal to adjust centre-assessed grades even where there is no evidence of inaccuracy is not well-founded</u>

The consultation document proposes to reject the approach of 'putting more weight on centre assessment grades'. This approach is described (p28) as one which assumes that 'these [i.e. centre assessment grades] are correct unless there is statistical evidence to the contrary'.

In rejecting this approach, therefore, the document is saying that centre assessed grades will not be assumed to be correct *even if there is no statistical evidence to the contrary*. This simply cannot be right. If there is no statistical evidence to suggest that centre assessment grades are inaccurate, then Ofqual and the Awarding Bodies can have no evidence at all to suggest that they are inaccurate, because they are taking no account of any other evidence (such as student work). The best available evidence is then that submitted by the centre, based on direct knowledge and experience of the student. Second guessing this appears to have no justification if there is no statistical justification.

We feel strongly that if there is no statistical evidence to the contrary, and in the absence of Awarding Bodies considering any evidence of student achievement, centre assessed grades should be accepted.

<u>Ofqual and the Awarding Bodies should take steps to strengthen centre assessed grading</u>

We can understand the concern that gradings received from centres may be substantially too lenient and the concern that different schools may apply different standards. Given these concerns, Ofqual and the Awarding Bodies should urgently take steps to strengthen the quality assurance at school level of grades being provided.

For example, it is perfectly reasonable to expect centres to take account of statistical data in their judgements and for Ofqual to specify how this is to be done (and to provide guidance and online training on this). It is also reasonable to ask centres to declare that they have done so; to specify any subjects where in the centre's view the attainment of pupils falls outside the range of achievement that would normally be expected (based on centre prior year performance and the prior attainment of students) and to provide an explanation of this; schools can also reasonably be asked to retain (and certify that they have retained) evidence supporting any such judgement.

Ofqual and the Awarding Bodies should have the right to scrutinise the centre-level evidence which supports the judgements centres have made, where these are out of line with prior years. Schools should also be expected to retain any in-year tracking data which has been used as management information. The chair of governors or other governance authority should be asked to certify that grades reported are consistent with information governors have seen over the year. It should be made clear that deliberate mis-reporting of grades is maladministration and that the usual maladministration processes and sanctions will be applied.

<u>'Statistically significant' variation from modelled grades is not a sufficient basis for altering teacher assessment grades</u>

Ofqual and the Awarding Bodies face an unenviable situation. In any normal year, the national pattern of year-on-year changes in results at centre level would follow a reasonably predictable national distribution. This will include changes at centre level which could not have been predicted with reasonable accuracy by historic data – and as in any such distribution, there will be outliers. This year, however, if centre submissions imply a national pattern of results which shows very substantial upward drift, then in some centres at least, the standard will have been misapplied and there will need to be some standardisation.

We see no way of generating a statistical model for doing this which can discriminate between genuine outliers (centres where there has been a significant change in standards achieved) on the one hand and centres which have significantly misapplied the standard on the other. Equally, it seems unlikely that centres which would have suffered a very sharp decline in results (i.e. will be

'negative outliers') will put forward results reflecting this. Again, there seems no way of identifying which these centres are.

Ofqual will need to take great care in determining what would be sufficient statistical evidence to change a school's gradings. While it is tempting to use basic tests of 'significance' as the grounds for making change, this needs great care and should be resisted in simple form. For example, if it were possible to create a model which predicted with '95% confidence' that the correct proportion of 4+ grades lay within a certain range, then we would expect 5% of schools' results to lie outside their modelled range. Simply correcting results to within the modelled range could therefore introduce substantial error to the results of some children in schools which had graded students perfectly accurately.

In this context, we think that standardisation should adopt a 'precautionary principle' of seeking to avoid unfair harm as far as possible. In particular, we think that it is preferable not to force conformity to a pre-determined statistical model and to take as broad a view as possible of statistical and other evidence. We set out some possible principles for this below.

<u>Ofqual must be open about the impact that the standardisation model would have had on prior year results, including the number of candidates who would have had grades reduced</u>

Any statistical model to be used for standardisation must be robust and tested on multiple prior year results. Ofqual and the Awarding Bodies should be open as to the level of error that would have been introduced had such a model been applied in prior years to actual examination results.

If the application of the standardisation model to be used in 2020 to the distribution of results actually achieved in 2019 and 2018 would have led to changes to many thousands of student results in large entry subjects in those years, then its validity is clearly questionable. Should it be the case that Ofqual and the Awarding Bodies wish to go ahead with a standardisation model with that characteristic, they will then need to explain transparently why the approach chosen is still to be used in 2020.

<u>Ofqual should be prepared to use all available evidence in standardisation and be prepared to use different models in different subjects and in different qualifications</u>

Any statistical approach used for standardisation should use all available statistical evidence to support decisions. Ofqual should be prepared to use different statistical models in support of awarding in different subjects. For example, at A level, there will be some subjects where prior attainment in that subject at GCSE is the best guide to A level performance. In other subjects, it may be that performance in a different GCSE or in a basket of GCSEs is the best guide to A level performance. Ofqual should use in each subject the model that (on the basis of testing on prior year results) best supports the fairest judgements in that subject.

Ofqual should also be prepared to apply different approaches at GCSE and A level. If a statistical approach shows strong predictive power and reliability at one stage but not the other, Ofqual should be prepared to use it at one stage but not the other.

<u>Ofqual should use several years' results and consider the range and variation in these results, not only measures of average results</u>

A centre's proposed grade distribution in a subject should not be changed only because it is out of line with the previous year's (or any single prior year's) results. Ofqual should consider a range of prior years' performance. However, it is insufficient to consider average performance over time –

the range and variance of results is more important than the average. If a centre's distribution of grades in a subject is not significantly different from that which would be expected given performance in at least one of the last three years, Ofqual should not in general change the centre assessment grades.

Ofqual should consider the range of performance information across different subjects in a centre and use evidence about the relationships between performance in different subjects

Ofqual should consider in depth the historic evidence about large changes in performance in a centre in different subjects and at different qualification levels and use this in building its models. Ofqual should consider whether it is possible to gain centre-level assurance from this about the reliability or otherwise of judgements in particular centres. For example: if it is very unusual to see substantial changes in performance in several large entry GCSEs in the same year, then this may be a more concerning pattern of results than in a centre where the vast majority of results are in line with expectations but one shows marked change.

Ofqual should consider in general whether it is possible to improve models or gain improved assurance through looking simultaneously at multiple subjects and levels in one centre. For example, is it possible to establish a level of assurance about the accuracy of judgements made in a centre, such that there would be confidence to accept a significant change in the pattern of grades in one subject? Or alternatively, are there certain patterns of results across multiple subjects, GCSE and A level which are extremely unlikely and give rise to reduced confidence in a centre's assessments?

Ofqual should adopt a process in which its approach to moderation can reflect the actual pattern of grades submitted by centres

Ofqual should consider remaining flexible in its approach to awarding until it has gathered empirical evidence from centres of their proposed assessment grades. If – in the best case – centres can do a good job of assigning grades to candidates in a subject such that Ofqual can see standards being broadly maintained in most subjects, it could adopt a relatively soft approach to standardisation. This might mean that teacher assessment grades can be accepted except in cases where there is strong evidence of inaccuracy.

If – in the worst case – centre assessed results in a subject are vastly out of line with any acceptable distribution in all subjects, Ofqual would need to adopt a different approach. It seems possible that the extent to which there are problems in grades submitted by centres may differ markedly between subjects. If it is desirable on the basis of assessed grades from schools to adopt different approaches to different subjects, Ofqual should be unafraid to do this.

Ofqual should re-consider the case for examining further evidence from some centres where there is a significant change in results

Finally, Ofqual should consider whether there is a case for examining further evidence from centres where there are significant changes in the results achieved. Statistical standardisation approaches can certainly do a good job of flagging apparently anomalous or unexpected results. However, they cannot tell us whether these apparently anomalous results are genuine outliers arising from the correct application of the process (of which there will be some in any year) or examples of inaccurate centre assessment.

It may be that Ofqual could consider such additional evidence in a limited category of centres. These could include centres which do not have any previous results (either at all or in a particular subject);

those where a significant proportion of students do not have prior attainment in tests or qualifications to be used in the model; those which have substantially changed entry patterns (including major changes to mix of qualifications, exam boards or moving from 'international' variant GCSEs or A levels); those where entry numbers in many or all subjects are very small; those which have had very significant changes in cohort size or composition; and those which have had recent changes in leadership and management.

We understand the logistical challenges this may present. How much examination of evidence from centres is possible may depend on the national distribution of grades proposed by centres. However, it is not unreasonable in principle for Awarding Bodies to be asked to assess the quality of evidence centres have for the grades awarded in cases where the distribution of proposed grades appears unusual. Ofqual could specify what is to count as stronger or weaker evidence and what strength of evidence would be required in order to substantiate outlying judgements.

Alternatively, or additionally, Ofqual could consider appeals from a limited range of centres against the standardised grade distribution in the centre. We agree that appeals from candidates which are essentially appeals against a candidate's position in the school's rank order cannot reasonably be considered. We also accept that the grounds for appeal from centres against standardisation decisions significantly affecting their proposed grade distribution need to be limited in scope to some extent. However, there seems merit in a process in which evidence of injustice arising from the application of a standardisation model can be heard and properly considered.

Conclusion

Ofqual and the Awarding Bodies have an unenviably difficult task in an unavoidable situation not of their making. We are concerned that the approach proposed in the Ofqual consultation document does not recognise enough that the harm done by under-grading candidates in this situation exceeds the harm done by over-grading candidates and that therefore there needs to be some relaxation of the standard; that Ofqual may be over-relying on a rigid statistical standardisation approach, when no statistical model can eliminate differences in standards between centres; that it would be a mistake not to assume that teacher grades are correct in the absence of evidence to the contrary; and that a rigid statistical model which over-writes teacher grades in this context would itself introduce unfairness.

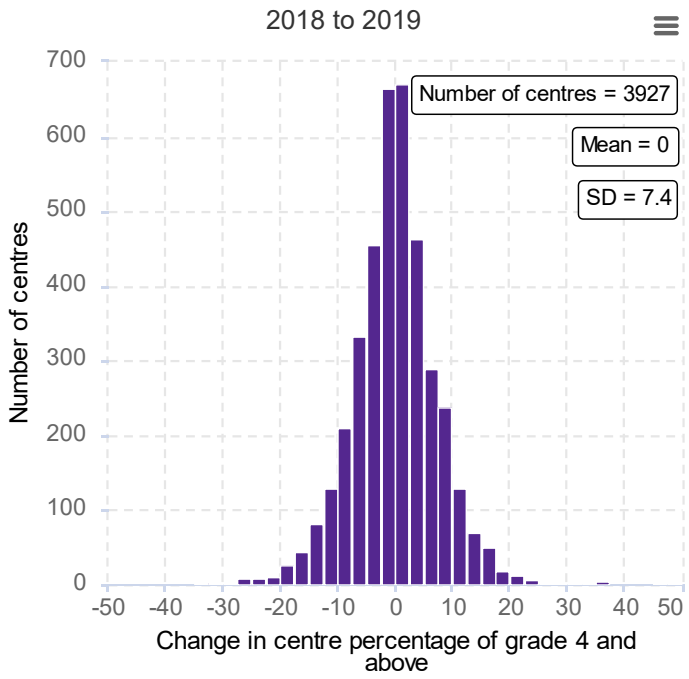
We do not pretend that the suggestions put forward in this paper can secure complete fairness, because this is not achievable in the current situation. Nor would we necessarily recommend the approach we advocate here in a normal year. But this is not a normal year and our priority must be to protect students who have already experienced severe disruption in a vitally important year of their education.

Jon Coles

24/4/2020

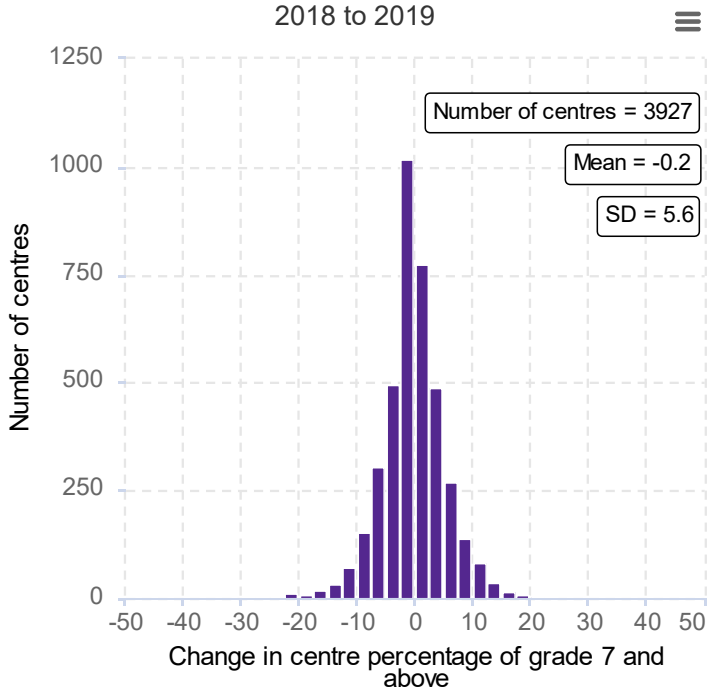
Annex: Examples of variability in achievement between centres year on year

Change in proportion of students achieving grade 4+ in English Language. (Centres of 25+ students.)



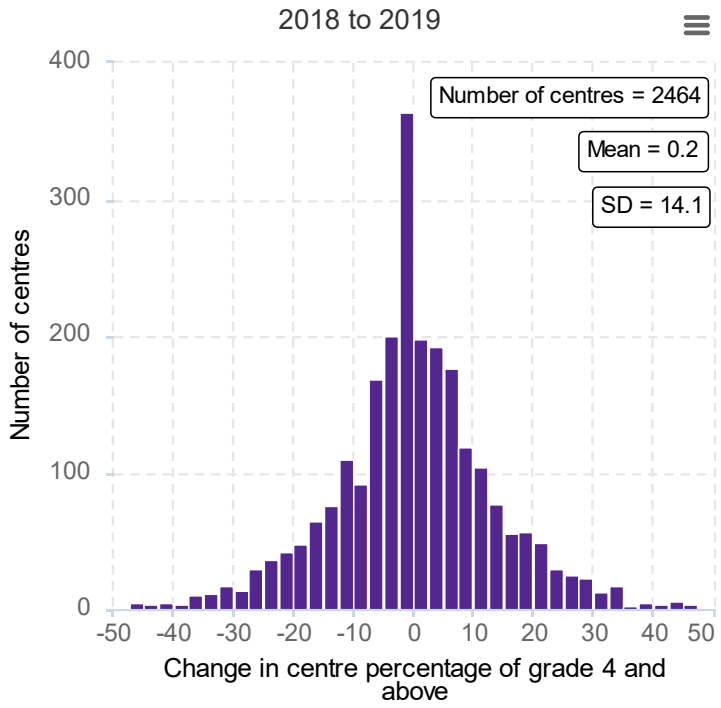
In 1335 centres of 3927 the change in proportion is no more than +/- 2.5%age pts (i.e. in 2592 centres, the change from year to year is greater than 2.5%age points). In Maths, the number is similar to that in English Language at 1332. The proportion in centres of 100+ students is similar.

Change in proportion of students achieving grade 7+ in English Language. (Centres of 25+ students.)



1797 centres within +/-2.5%

Change in proportion of students achieving grade 4 in Art & Design. (Centres of 25+ students.)



In 562 of 2464 centres, the difference in proportion achieving grade 4+ between 2018 and 2019 is no more than +/-2.5%

A level mathematics – change in proportion achieving grade A between 2018 and 2019