Centre for Education Research and Practice

Setting the grade standards in the first year of the new GCSEs

William Pointer

Summary

Reformed GCSEs in English, English Literature and Mathematics are being introduced for first teaching from September 2015 with the first examinations in summer 2017. Other subjects are being reformed to start the following year with first examinations in summer 2018. The new specifications will be assessed linearly, will have revised subject content and will have a numerical 9-point grade scale.

This paper looks at the results of simulations that were carried out to inform how the new grading scale for GCSEs will work. It discusses the pitfalls associated with various ways of implementing the new grade scale, highlights potential problems that could arise and evaluates the final decisions made by Ofqual. The paper focuses specifically on issues relating to the transition year and not subsequent years.

Ofqual has decided that the new grading scale should have three reference points: the A/B boundary will be statistically aligned to the 7/6 boundary, the C/D boundary will be mapped to the 4/3 boundary and the G/U boundary will be mapped to the 1/U boundary. This will aid teachers in the transition to the new grading scale and will also aid employers and further education establishments to make more meaningful comparisons between candidates from different years. If possible, pre-results statistical screening will be used to ensure comparability between awarding organisations at all grades, not just those that have been statistically aligned, by means of predictions based on mean GCSE outcomes.

The above decisions seem logical based on the modelling which has been carried out. However, the decision to set grade 9 so that the percentage of candidates achieving this top grade is 20% of the cumulative percentage obtaining grade 7 is more controversial. It risks having narrower grade boundaries for the top grades and also fails to take into account differences in the ability of the candidates taking different subjects.

For tiered subjects Ofqual has said that test equating will be used at grades 4 and 5 to inform boundary setting; ensuring standards are equated between tiers. However, there is perhaps also a place for examiner judgement in a confirmatory capacity.

An increase in the number of grades is likely to have a detrimental effect on classification accuracy, i.e. more candidates are likely to be awarded an incorrect grade. This can be mitigated through assessment design but may have to be accepted as a consequence of the desire to have more differentiation among higher-performing candidates.

Introduction

Reformed GCSEs in English, English Literature and Mathematics are being introduced for first teaching from September 2015 with the first examinations in summer 2017. A host of other subjects are being reformed to start the following year with the first examinations in summer 2018. The new specifications will be assessed linearly, will have revised subject content and will have a numerical 9-point grade scale. The grades will run from 1 to 9. Grade 9 has been

AQA Education (AQA) is a registered charity (number 1073334) and a company limited by guarantee registered in England and Wales (number 3644723). Our registered address is AQA, Devas Street, Manchester M15 6EX.

set as the highest grade rather than grade 1 (as was the case with CSEs) to enable the future addition of extra grades giving further discrimination at the top end of the ability range should the need arise. In his policy steer, the then Secretary of State for Education, said:

> *"Any changes should apply across all subjects, and should differentiate performance more clearly, particularly at the top end."*
>
> *(Gove, 2013)*

In April 2014 Ofqual launched a consultation on the new grade standards for GCSEs (Ofqual, 2014a). Ofqual's primary aims with regard to the new grading scale are to:

- ensure candidates are not disadvantaged in the first year of the new GCSEs
- have reference points to enable comparisons between the old and new scales
- increase discrimination above grade C.

In the consultation Ofqual discussed whether, in place of the current method, using a criterion or norm-referenced approach to standard setting would be appropriate. However, Ofqual concluded that these approaches would not be suitable and proposed that standard setting would continue along the established methodology using statistics to achieve comparable outcomes. This approach will help to ensure candidates are not disadvantaged in the first year of the new GCSEs. However, this is unlikely to be straightforward as the reforms could see large-scale shifts in entry patterns, which could affect value-added scores and invalidate the predictions[1]. Therefore additional information, including the use of stable common centres, may be needed to award the subjects reliably in the first year.

As part of the consultation on the new grades, the awarding organisations undertook some modelling with empirical and simulated data. The government's policy intention was for the new GCSEs to be more demanding than the current ones. Ofqual stated in its consultation document that "the breadth and depth of the subject content of the new GCSEs will typically be more demanding than that for current GCSEs" (Ofqual, 2014a). Therefore the mark distributions could vary wildly from what has been seen in the past. AQA carried out simulations using R (R Core Team, 2014) to understand how different mark distributions might translate into the grades that would be awarded for the new GCSEs. Mark distributions were generated from a skew normal distribution (Azzalini, 2014) with a given mean, standard deviation and skew. The current A*-G grade boundaries were then fitted to the mark distributions based on given subject cumulative percentages at the judgemental grades A, C and F from 2013. The new grades were then fitted to the mark distribution for the various different models that were under consideration. Further details of the simulations can be found in Appendices 1-3. The model numbering in this paper may seem arcane and non-intuitive; this is because the original model labelling has been retained because it cross-references what was done by JCQ for Ofqual (see Appendix 3 for full details of the models).

Ofqual released its decisions for the new grading scale on 12 September, 2014 (Ofqual, 2014b). These decisions were to a large extent informed by the work of the awarding organisations. This paper attempts to link the results of the simulations with the decisions that Ofqual made for how the new grading scale will be constructed. It also considers the wider implications of these

[1] For example, in summer 2014 predictions for GCSE Science were based on 16-year-olds instead of 15-year-olds because it was felt that a prediction based on 15-year-olds would not be valid. This was in part due to a large, non-random reduction in the number of 15-year-olds entered which was precipitated by the government altering the rules governing performance tables so that only a candidate's first result would count.

decisions on the communication of relative standards, the accuracy of grading and the mechanisms and entry policies for tiered subjects.

In subsequent series, Ofqual plans to make use of a new national reference test to help inform standard setting. The aim of this is to allow outcomes to rise or fall if there is an increase/decrease in the ability of the candidates, something that is not possible under a comparable outcomes approach given that adjustments are made to nullify any overall rise or fall in the mean Key Stage 2 (KS2) scores used to measure prior attainment. The national reference test, and how it will be applied, is currently under review. For this reason, and because it is not relevant to the setting of standards in the first year, it is not discussed any further in this paper.

A new grading scale

Reference points

The mid-grades

One of Ofqual's aims was to ensure that there was some degree of comparability between the current grade scale and the new one. To this end, Ofqual proposed from the outset that the grade C/D boundary should be statistically aligned to the grade 4/3 boundary. This will be achieved by setting the grade 4/3 boundary so that the cumulative percentage of candidates achieving a grade 4 is as close as possible to the cumulative percentage suggested by a prediction for grade C (where the prediction is based on the relationship between KS2 scores and outcomes from all awarding organisations from a given reference year). This prediction will allow standards to be aligned across years and across awarding organisations at what is currently a key grade. Once the alignment between the 4/3 boundary and the C/D boundary had been decided, an agreement was needed as to whether any other boundaries would be statistically aligned so that the standard at a current boundary would be mapped across to one of the new grades.

The first phase of the simulations included two models (models 3a and 3b, see Appendix 3 for full model details) where there was only one reference point, grade C mapped to grade 4. All other boundaries were set by linear interpolation so that they had equal grade widths[2], model 3a, or so that there was an equal percentage of candidates at each grade (including the percentage of candidates who were ungraded), model 3b.

The effect of these models on the grade distributions is displayed graphically in Figure 1. The left hand plot shows the percentage of candidates at each grade whilst the right hand plot shows the width of each grade in terms of marks. The data used to derive the plots is based on a simulated mark distribution with a mean of 50% of the maximum mark, a standard deviation of 16% of the maximum mark and no skew. The current judgemental grade boundaries were set so that the cumulative percentages matched the results from all subjects from June 2013 (JCQ, 2013)[3]. The other boundaries were set arithmetically using the established methodology as specified in Ofqual's Code of Practice (Ofqual, 2011). The colour bands denote the different grades and the dotted lines that run through the plots denote the grade boundaries under the current A*-G grading scheme. By comparing the dotted lines on the plots for model 3a and 3b

[2] The boundary widths of all grades above 4 were the same; similarly, the grades below grade 4 all had the same width.

[3] Grade A = 21.3%, Grade C = 68.1%, Grade F = 96.8%

with the colour bands, it is possible to understand the relative impact of the two new grading schemes.

Model 3a, where boundaries were set to have equal widths, fails to spread candidates out at the populous grades. In fact it splits the A* candidates into three grades whilst grades A, B and C are not spread out as they are split between grades 4, 5 and 6. Model 3b, where boundaries were set so that each boundary has an equal percentage, sees a large increase in the percentage of candidates failing (i.e. gaining a U). There is also no extra differentiation at the top end as A and A* candidates will still be spread out over just two grades. Looking at the plot on the right we can see also that the grade widths for the middle grades, in particular, are especially narrow, so there would be an increased risk of grades being misclassified.

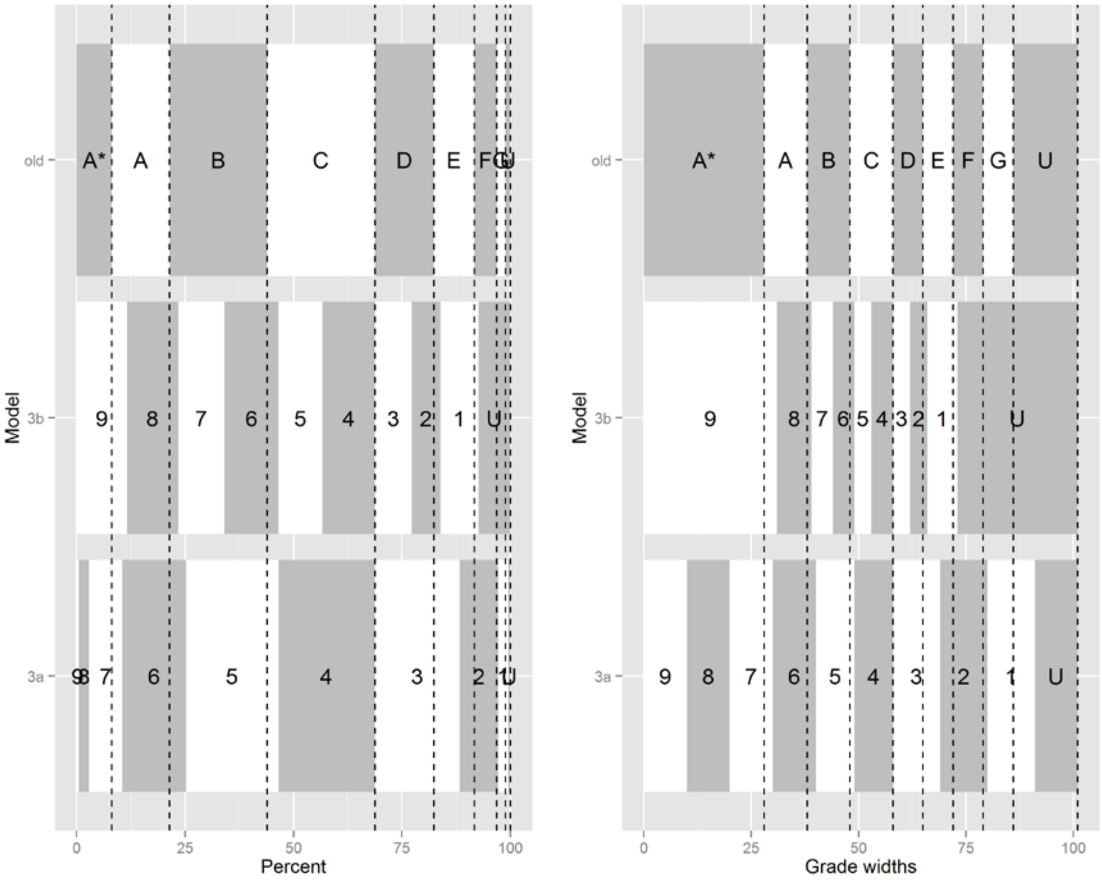


Figure 1 **Percentage of candidates at each grade and the width of the boundaries where there is only one fixed reference point.**

Therefore it seemed there was a need for more reference points. Other than grade C, the judgemental boundaries for GCSE are grades A and F. These provided the logical basis points for any further mappings as standards across time and awarding organisations are currently maintained at these boundaries.

The top grades

Mapping the bottom of grade A to the bottom of grade 7 was the logical reference point as it would split grades B and C into grades 4, 5 and 6 and grades A and A* would be spread out over grades 7, 8 and 9. The relative grade widths can be seen in Figure 2, for instance it can be seen that the grade 5/4 boundary would be two-thirds of the way between grades C and B.

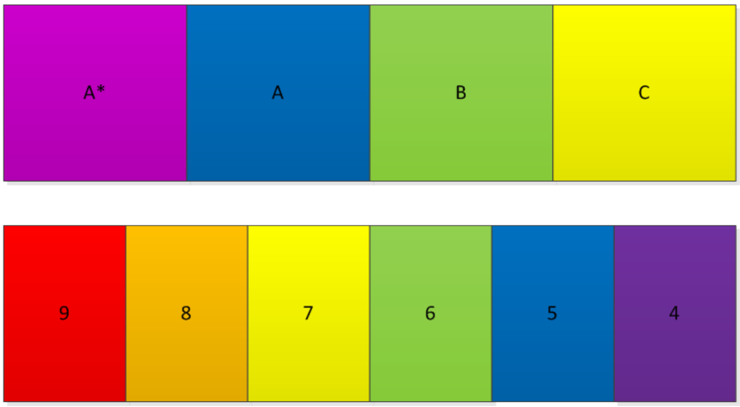


Figure 2 Relative grade widths if grade A was mapped to grade 7.

Some consideration was given to setting grade 7 to be four-fifths of the way between grades C and A or setting grade A to be grade 6. These possibilities were not extensively modelled, nor have they been reported in this paper, as they were not considered viable alternatives; adding little or no extra differentiation in the important and overcrowded middle grades B and C. Ofqual indicated in its announcement of 12 September that grade A would indeed be mapped to grade 7.

No child left behind

The new 9-1 grading scale has been designed to offer more discrimination at the top end, with six grades above what is currently grade C in place of four. This comes at the expense of fewer grades below what is currently grade C. In his policy steer in February 2013 the then Secretary of State for Education said:

> *"The reformed GCSEs should remain universal qualifications of about the same size as they are currently, and accessible, with good teaching, to the same proportion of pupils as currently sits GCSE exams at the end of Key Stage 4."*
>
> *(Gove, 2013)*

To ensure that there is control over the percentage of candidates who are unclassified, so that the qualifications remain accessible to the same number of candidates, there needs to be a reference point from the current scale at the bottom of the range. Three possible models were considered (see Appendix 3, for full model details). Figure 3 shows the relative grade widths under the three different models considered.

The first two models considered mapping the F/G boundary across to the new scale. Grade F is currently a judgemental grade so is a logical choice to be a reference point as standards are currently maintained at this boundary. The first model mapped grade F to grade 2 (model 1). In this case there would be two grades to replace grades D, E and F, with each of these grades one and a half times the width of the current grades. In 2013, 21.5% of candidates obtained a grade D in English (JCQ, 2013). If this model was adopted, the width of grade 3 would be wider than the current grade D so over a quarter of all candidates would achieve this grade; thus this model would not discriminate very efficiently. Also, the grade width for grade 2 would be wider than grade F so grade 1 would be wider than grade G (as these are arithmetic boundaries[4]) and

[4] Grade G is set so that it spans the same number of marks as grade F, and under this model grade 1 is set so that it spans the same number of marks as grade 2.

consequently there would be fewer candidates being unclassified. This would contradict the notion that the new GCSEs are to be more demanding.

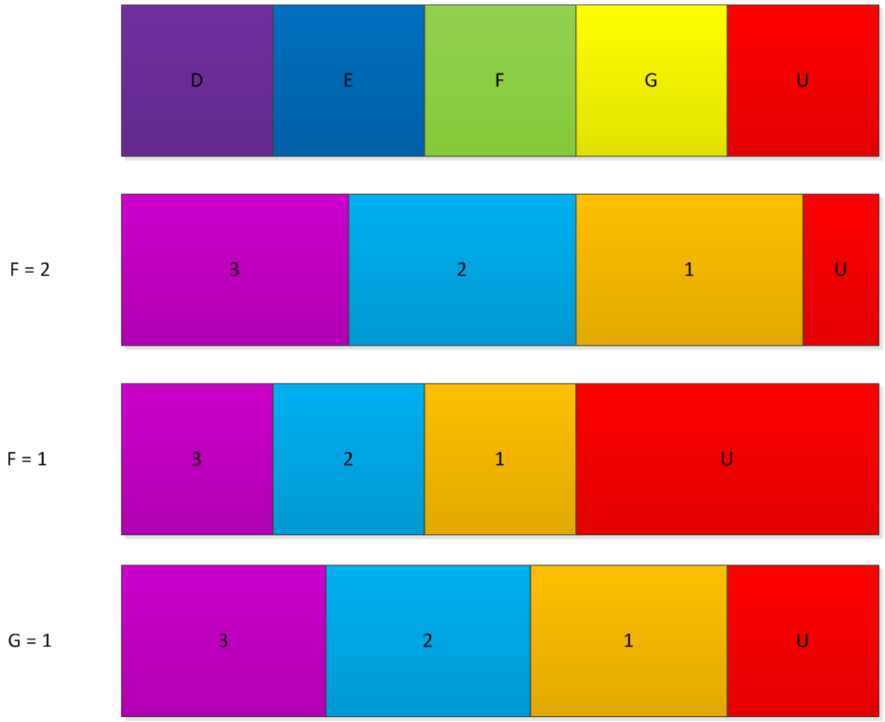


Figure 3 Relative grade widths if grade F was mapped to grade 2.

The second model mapped grade F to grade 1 (model 2). Under this model the existing grades D, E and F would simply become 3, 2 and 1 respectively. The notable downside with this method is that anyone who would have obtained a grade G will now be unclassified. Table 1 shows the percentages of candidates who achieved grades F, G and U in 2013. It can be seen that grade G tends to be a sparsely populated grade – in the case of English Literature, fewer than 1% of candidates obtained a grade G. However, nearly 5% of candidates obtained a grade G in Mathematics so the difference would be much more noticeable here as the fail rate would rise from 2.7% to 7.4%. This would contradict the notion that the reformed GCSEs are to be accessible to the same proportion of candidates as the current GCSEs.

Table 1 Percentages of candidates at grades F, G and U in 2013.

Subject	F	G	U
All subjects	4.1	2.0	1.2
Mathematics	6.9	4.7	2.7
English	3.8	1.2	0.7
English Literature	2.2	0.9	0.8
Geography	4.3	2.2	0.8
Religious Studies	4.4	2.7	1.7

Source: JCQ, 2013

If the proportion of candidates failing is to remain unchanged then grade G must be mapped to grade 1; this was the third model considered (model 6). One problem with mapping grade G to grade 1 is that the standard of grade G is somewhat arbitrary, as the boundary is calculated arithmetically and so depends on the mark distribution. Therefore the standard can vary from year to year and between awarding organisations. However, a prediction can be generated based on the outcomes from all awarding organisations in the reference year to enable common standards to be set across the awarding organisations.

This approach would mean that four grades were being condensed into three, so the grade widths would be wider than they currently are. It would increase the percentage of candidates obtaining a grade 3 which, as already discussed, is a populous grade.

Given these considerations, Ofqual decided that grade G should be mapped to grade 1 (model 6). With a comparable outcomes approach in place in the first year this ensures that there will not be a lessening in the pool of candidates for whom GCSEs are attainable. The possible downside is that grade 3 may be awarded to a large percentage of candidates so the grading scale will not discriminate very well at this section of the ability range. However, Ofqual noted that, in response to the consultation, many teachers said that a grade G represents a real achievement for some candidates. The intended approach ensures that these candidates will not be disenfranchised.

Intervening boundaries and common standards

With the decision in place that grades 1, 4 and 7 should be statistically aligned (using comparable outcomes) with the current grading system, two main methods of setting the intervening boundaries were modelled. The first used equal grade widths and the second used an equal percentage of candidates at each grade. The current grading scale uses equal grade widths to set intervening boundaries.

Any potential model is subject to variations in the mark distribution; if the standard deviation is low then the boundaries will be compressed. The major difference between a fixed width and a fixed percentage approach is that a fixed percentage approach risks having very narrow boundaries so increases the frequency of grade misclassifications, whereas a fixed width approach risks having an unbalanced distribution of candidates across grades so that some grades may have few/no candidates and some may have many.

Mark distributions do not tend to be uniform so why would we expect the grade distributions to be uniform? In fact, mark distributions tend to be roughly normally distributed so setting boundaries to have equal percentages risks having extremely narrow grade boundaries for the middle grades, leading to an increase in grade misclassifications in this section of the ability range. The modelling that was conducted using an equal percentage approach confirmed this hypothesis.

Informed by this modelling, Ofqual decided that the intervening boundaries (2, 3, 5 and 6) should be set by interpolation (grades 8 and 9 are discussed separately in the next section). This decision, however, introduces new problems to the maintenance of standards. While boundaries based on the reference points will be set using statistical predictions, so standards can be maintained across time and awarding organisations, for the other boundaries there is no certainty that standards will be aligned. This was always the case with the old grading scheme but is of particular importance with the new grading scheme because, in due course, grade 5 will become a key threshold.

Each year, following the release of results, the awarding organisations carry out a statistical screening of outcomes. The process is based upon using mean GCSE scores as a measure of candidate ability to check that outcomes are comparable between the awarding organisations (and over time). The outcomes of the statistical screening are then used to inform standard setting in the subsequent year. A trial is planned this autumn to see if it would be possible to carry out the statistical screening *before* results are published so that changes could be made to address differences in standards straight away. (For more detail on the pre-results statistical screening trial see Eason, 2014.) If such a process proves viable it is possible that it could be used to ensure standards are comparable at *all* boundaries, not just the reference points, by amending boundaries from their arithmetic positions based on the screening outcomes.

In the short term, with the staged introduction of the new GCSEs, a major issue affecting the use of the pre-results statistical screening process is the fact that two different grading scales will be in use. Therefore mean GCSE score might have to be restricted to the legacy specifications that still use A*-G. It is not known how this might affect the outcomes of the screening process. Even if pre-results statistical screening is deemed viable in the current context, a more detailed investigation of whether it would be a suitable method in the first year of the new GCSEs should be undertaken before being implemented.

An alternative solution, if the pre-results statistical screening proves to be unfeasible, would be for each awarding organisation to re-grade the reference year specifications using the new grading scale. These results would then be used as the basis of predictions for the first year of the new GCSEs. The boundaries that are calculated arithmetically would be checked against the prediction and adjusted if they were outside of a given tolerance, say 2% (this mirrors the approach currently used for the A* boundary). This process could be carried out at all intervening boundaries or just the boundaries deemed to be key for the future: probably grades 5 and 8. However, the vagaries of current mark distributions would affect the outcomes.

Ofqual has expressed a desire to use pre-results statistical screening to ensure comparability between awarding organisations at all boundaries (Ofqual, 2014b). As long as it is possible to implement this process without compromising the timescales for releasing results then it should ensure that results are aligned at each boundary across awarding organisations. In subsequent years it would be possible that all boundaries, not just the key boundaries, could be set using predictions. Although the grade boundary widths could be uneven, standards would be maintained at each boundary rather than just the key boundaries, which is surely an improvement on the current state of play.

The grade 9 quandary

Setting boundaries at the extremes of the mark distribution can prove problematic. In the current system, a prediction is generated for A* and boundaries are moved from the arithmetic position if the outcome would be outside of a 2% tolerance from the prediction. To determine a method for setting the top grades in the new grading scheme, simulations focused on this area.

The methods that were modelled were to set the grade 9 boundary (see Appendix 3, for full model details):

- arithmetically, so that the grade widths for grades 7 and 8 were the same as for grade 6 (model 2a)
- by interpolating between the grade 7 boundary mark and the maximum mark (model 2c)
- such that the percentage of candidates at grade 9 was half the percentage of candidates who would have obtained an A* (model 2d)
- such that the percentage of candidates at grade 9 was 20% of the cumulative percentage of candidates at grade 7 (model 2e).

One of the aims of the new grading scale is to increase discrimination at the top end. Depending on the method chosen to calculate grade 9 and the vagaries of the mark distribution, it would be possible that very few, perhaps no, candidates would achieve a grade 9; in this case the A and A* candidates would be split between grades 7 and 8 and discrimination would be no greater than it is now. Alternatively there could be more candidates achieving a grade 9 than currently get A*. However, this might be acceptable as the grade A candidates would be spread out over two or more grades (7, 8 and part of 9), increasing discrimination.

Interpolation method – model 2a

Interpolating between the grade 7 boundary and the maximum mark would mean setting grades 8 and 9 using the same methodology as for the other arithmetic boundaries. This method was ruled out in the early phase of modelling as it quickly became apparent that there could be very few or no grade 9s. Figure 4 shows a histogram of a mark distribution where the marks are normally distributed, the maximum mark is 100, the mean is 50 and the standard deviation is 16. The positions of the grade boundaries are overlaid.

Fewer than 1% of candidates achieve a grade 9. In fact, there are fewer candidates obtaining a grade 8 or above than would have achieved an A*. This outcome was seen in numerous simulations as well as in the modelling of empirical data so this method was not pursued any further. The problem occurs if the grade 7 mark is quite low as a percentage of the total mark. It could often occur in the new GCSEs since the demand is being raised but the percentage of candidates achieving a grade 7 is being maintained. If the demand is raised too high it is possible that the boundary marks may be uncomfortably low as there will be dead marks at the top of the mark distribution.

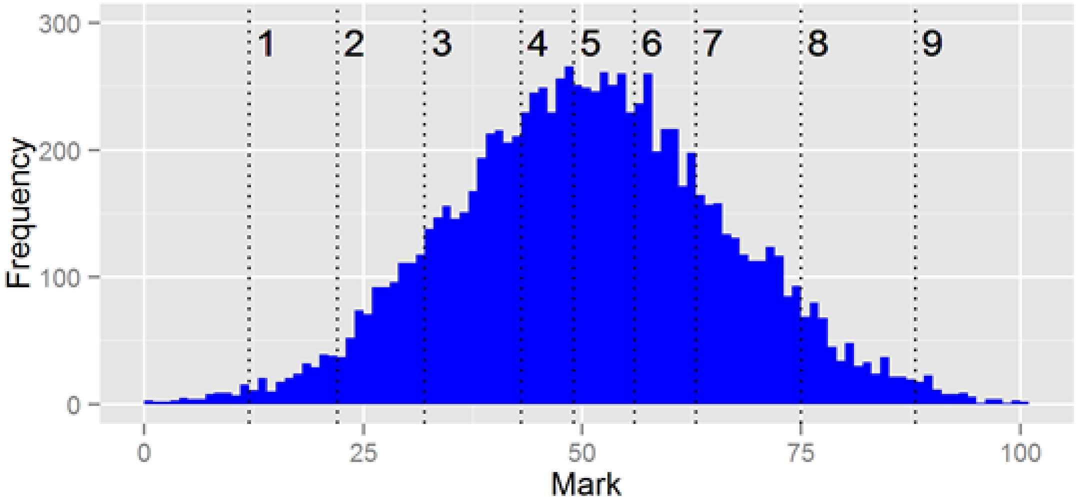


Figure 4 **Histogram of a simulated mark distribution with grade information where the top grades are interpolated between grade 7 and the maximum mark.**

For the remaining three models under consideration, simulations were run using the cumulative percentages for grades A, C and F from five different subjects (see Table 5 in Appendix 1 for details). The mean mark varied from 40% to 60% of the maximum, the standard deviation varied from 12% to 20% of the maximum and the skew varied from -3 to 3. The histogram of an exemplar mark distribution with skew of -3 is given in Appendix 4. In total 1,925 simulations were run for each model in this phase of the simulations.

Arithmetic method – model 2c

The arithmetic method is where grades 8 and 9 are set so that the grade widths above grade 7 are the same as they are below. This leads to a linear relationship between marks and grades from grade 4 up to grade 9. Therefore the risk of grade misclassification at grades 8 and 9 will be no higher than for grades 4 to 7.

The ratio of grade 9s to A*s was used as a metric to evaluate the extent to which other models effectively discriminate at the top end. Figure 5 shows the percentages of A* candidates who would obtain a grade 9 in the various scenarios that were simulated in this phase of the modelling. The darkness of the points is based on the density. It can be seen that, using interpolation (model 2a), there might be no or very few grade 9s being awarded – in fact in almost half of the simulations the ratio of grade 9s to A*s was less than 10%.

Within the confines of the simulations, the arithmetic model (2c) looks promising as the ratio of grade 9s to A*s varied between 24.0% and 82.6% but was generally above 50%, suggesting that there would be fewer grade 9s than A*s but that grade 9 would not be a 'desolate' grade. It also suggests that there would be more discrimination in the top grades.

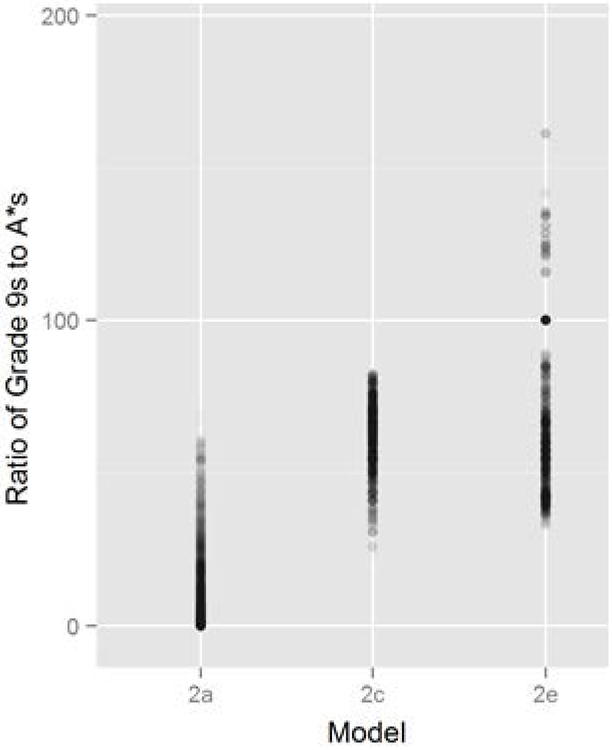


Figure 5 Ratio of grade 9s to A*s depending on how grade 9 is set.

However, if the mark distribution is positively skewed the arithmetic method could give rise to more grade 9s than grade 8s. This situation occurred in 27% of the simulations for this model whereas it did not happen at all when using the 20% rule (see model 2e below). It must be noted that not all of the simulated scenarios are equally likely to occur because the probability of each possible simulation is not the same and is not (easily) quantifiable, so (despite the 27%) it could be very unlikely that there are more 9s than 8s. In any case, it could be a product of the candidature that there are more 9s, for instance in the classical subjects more candidates get an A* than an A in the current GCSEs.

In the modelling, when the grade 9 boundary was calculated arithmetically and the width between the maximum mark and the grade 7 boundary was less than three times the width between grade 6 and 7, then grades 8 and 9 were set by interpolating between grade 7 and the maximum mark. This was done to mimic the procedure used to calculate A*. Stringer (2014) argues that, in a linear world, there are no technical reasons for not setting the A* boundary as the maximum mark. So the grade 9 boundary could always be set arithmetically rather than by interpolation between 7 and the maximum mark, unless it was greater than the maximum in which case it would be set as the maximum.

*50% of A*s – model 2d*

In its consultation, Ofqual suggested setting the grade 9 boundary so that the percentage of candidates at this grade was half the percentage of candidates who would have obtained an A*.

The drawback with this method is that it would perpetuate any issues that currently exist with A* and it fails to take into account changes in the ability of the candidature, so could unfairly (dis)advantage a particular year group.

The 20% rule – model 2e

The final option considered was to set the grade 9 such that the percentage of candidates at grade 9 is 20% of the cumulative percentage of candidates obtaining grade 7. Grade 8 would then be set midway between grades 7 and 9. This is subsequently referred to as the 20% rule.

We can see the result of the simulations for the 20% rule in Figure 5. We can also work out the approximate percentages of candidates obtaining a grade 9 directly from the 2013 outcomes[5]. Table 2 shows (for certain subjects) the cumulative percentages of candidates obtaining grades A* and A and the cumulative percentages expected to get a grade 9 using the 20% rule. It can be seen that the ratio of grade 9s to A*s varies quite dramatically by subject, for instance in Science there would be more grade 9s than A*s but in Chemistry there would be half as many grade 9s as there are A*s. This is due to the differences in the ratio of A*s to As that are currently awarded.

Table 2 Cumulative percentages of candidates at grades A* and A and the number that would achieve a grade 9 if the 20% rule was used.

Subject	A* Cum %	A Cum %	9 Cum %	Ratio of 9s to A*s	Ratio of A*s to As
Mathematics	4.9	14.3	2.9	59.2	34.3
English	3.3	14.2	2.8	84.9	23.2
English Literature	5.5	22.8	4.6	83.6	24.1
Chemistry	16.6	42.2	8.4	50.6	39.3
Science	1.4	8.1	1.6	114.3	17.3

Inter-subject comparability is a complex topic; it is debatable whether it is even a realistic aim. However, comparability of cognate subjects, for example Science and the separate sciences, is perhaps more relevant. If there was a shift in the method of awarding the top grade it might be substantially easier to obtain a 9 in Science than in Chemistry. This could lead to a further reduction in the number of candidates taking the separate sciences, contradicting current government policy intentions.

The ratio of grades is not maintained at any other level. For example, looking at the ratio of As to Cs shows wide variation between subjects, as one would expect as the candidature is different between subjects. The new system will not generally look to address this imbalance so it seems peculiar to do it uniquely at grade 9.

If the ratio of grade 9s to grade 7s is set at 20%, we fail to acknowledge the differences in the profiles of candidatures for different subjects. It could be conceded that the ratios of A*s to As is currently unbalanced so the opportunity to redress the imbalance should be welcome. However, the 20% rule seems a particularly blunt instrument and the attempt to redress a supposed imbalance between subjects is only being done at grade 9, not at any others. There is also the risk of differences between awarding organisations as the ability profile of the candidature at the top end could vary, and as a consequence there could be diverging

[5] It is an approximate percentage because the actual percentage may vary depending on how many candidates are at each mark.

performance standards. This is particularly important at grade 8 as this could be a significant grade in the same way as A is currently.

The benefit of the 20% rule is that it ensures that some candidates will always get a grade 9, whilst ensuring that it is not overly populous. Exerting this control does mean that if the mark distribution is too compressed or negatively skewed there is the likelihood of narrow grade widths and an increased risk of awarding incorrect grades. If the arithmetic method was used in these cases, there could be very few grade 9s although this situation could be remedied by altering boundaries based on the statistical screening outcomes. In the simulations, the 20% rule resulted in narrower grade widths for English in over three quarters of the cases compared to the arithmetic method. On the other hand, subjects with a higher ratio of A*s to As, for example Religious Studies, had wider grade widths under the 20% rule than under the arithmetic rule.

Further deliberations on the new grading scale

A ‘good pass’, these days, is hard to find

In his policy steer in February 2013, the then Secretary of State for Education said:

> *“At the level of what is widely considered to be a pass (currently indicated by a grade C), there must be an increase in demand, to reflect that of high-performing jurisdictions. This is something we believe the vast majority of children with a good education should be able to achieve.”*
>
> *(Gove, 2013)*

Grade C is currently considered to be a good pass and is used in the performance tables compiled by the Department for Education (DfE). Since grade 4 is to be statistically aligned with grade C using a comparable outcomes approach, it cannot be seen to reflect an increase in demand. Therefore it seems likely that a grade 5 will be the new good pass. Across all subjects in 2013 the cumulative percentage achieving a good pass, i.e. a grade C, was 68%. By fixing grade 7 to grade A, grade 5 will fall two-thirds of the way between grades B and C so the good pass rate will be approximately 51%, a fall of 17%[6]. If this is used as the pass mark for performance tables, as seems likely, then it will have an impact on schools[7].

In terms of linking outcomes to international standards, Ofqual appears to have taken a light touch in that the grade 5 boundary will not be set to directly tie in with some notion of an international standard. A more direct method of linking grade 5 to an international standard could have led to conflict with standards at the other grades, for example the grade 6 boundary could have fallen absurdly close to the grade 5 boundary or possibly even below it. Instead, Ofqual sees grade 5 as naturally falling in a place that will reflect the performance of high performing jurisdictions. This notion is based on a report from the DfE that looked at PISA[8] results and concluded that the gap between the UK and the high-performing jurisdictions was between half and one grade (DfE, 2011).

[6] Using the rather simplistic notion that one third of those currently awarded a grade C will gain a grade 5.

[7] This will not affect the new Progress 8 or Attainment 8 measures but will impact the EBacc and English and Mathematics measures which are based on the percentage of candidates obtaining a good pass.

[8] Programme for International Student Assessment

However, simply raising the threshold for a good pass will not in itself improve performance standards or raise the UK's performance in PISA tests in the future; this will only happen if there are genuine changes in the performance of candidates brought about by improvements in teaching.

The DfE is not the only external stakeholder that will be looking to interpret the new grading scale. How will others (e.g. businesses and further education) compare candidates who have results on the A*-G scale with those who get results on the new numerical scale? That will be up to them, but having statistically equivalent points will be helpful. However, there is a potential misinterpretation that comes from statistically aligning some of the boundaries. Whilst it would be correct to say that the bottom of grade 4 is the same as the bottom of a grade C, this does not mean that grades 4 and C are directly comparable. A grade C could equate to a grade 4 or grade 5. A clear, coherent communication strategy from Ofqual and the awarding organisations is essential for the successful implementation of the new grading scale. It is encouraging to see that Ofqual has noted this in its statement accompanying the release of information on the new grading scale (http://ofqual.gov.uk/news/setting-standards-new-gcses-2017/).

Tiered subjects

Most of the new GCSEs will be untiered. However, Mathematics, the sciences and modern foreign languages[9] will retain tiering. Candidates will have to be entered for the same tier for all units. Higher tier candidates will be able to obtain grades 4 to 9, with an allowed grade 3[10]. Foundation tier candidates will be restricted to grades 1 to 5. Currently the higher tier is targeted at grades A*-D (with an allowed E) whilst grades C-G are possible on the foundation tier. With the bottom of grade 4 being statistically aligned to the bottom of grade C, the ability range for which the higher tier is appropriate will be reduced – candidates currently achieving grade D will be liable to fall off the bottom of the scale and be unclassified. In 2013, nearly 15% of candidates entered for the higher tier of AQA's linear Mathematics GCSE failed to achieve a grade C; these candidates would be in danger of getting a U.

Schools will need to be aware of the implications of this shift in the balance of the tiers so that they enter candidates for the appropriate tier. With many schools streaming candidates from the start of KS4 (or before), it is important that they are made aware of the implications before teaching of the new specifications commences in 2015.

The upshot of a large shift in entry patterns is that any tier-level predictions will be unreliable. The value-added rate is not the same for foundation tier and higher tier candidates so it will only be possible to have an aggregated subject-level prediction.

Currently AQA uses test equating when setting the standard for grade C in many tiered subjects. Whilst the subject level prediction drives the award, test equating is the main mechanism (supplemented by examiner judgement and tier-level unit predictions) to ensure that the standards are equivalent across the tiers. Test equating could be used in conjunction with the subject-level prediction to set the boundary for grade 4.

On untiered papers grade 5 is interpolated between grades 4 and 7. This method would be possible on the higher tier but not on the foundation tier. One possible method on the

[9] Tiering has been confirmed for French, German and Spanish. A decision has yet to be made for the smaller entry MFLs.

[10] The allowed grade 3 is set such that the boundary width for grade 3 is half the width of the grade 4 boundary.

foundation tier would be to set grade 5 to be as many marks above grade 4 as grade 3 is below. However there is no guarantee that the standards will be aligned at this key boundary if this (or any other arithmetic) process is employed.

Test equating was performed on the AQA linear Mathematics GCSE from summer 2012 and summer 2013. When the subjects were originally awarded, test equating was used to align the grade Cs; thus the equivalent grade 4 boundary marks were equated (to within a mark). Table 3 shows the test equated pairs of marks for grade 5 for 2013 and 2012. The highlighted marks are the arithmetically calculated boundary marks using interpolation/extrapolation. In both years the test equating suggested that the grade 5 boundaries would not be aligned and that the boundary mark on the foundation tier would be harsh in relation to the higher tier mark. Since it seems unwise to move the interpolated higher tier boundary, the foundation tier boundary must be moved. In the 2013 case it would need to be lowered by 8 or 9 marks from the arithmetic position to a mark of 140 or 141. This would reduce the grade 4 boundary width to just 15 marks (8.6% of the maximum mark) and the percentage of candidates achieving a grade 5 on the foundation tier would rise from 6.20% to 12.48%. A similar shift would be needed based on the data from summer 2012.

Table 3 Test equated marks for grade 5 AQA linear Mathematics GCSE from summer 2012 and summer 2013.

2012				2013			
Cum %	FT	HT	Cum %	Cum %	FT	HT	Cum %
8.88	138	67	76.19	13.54	140	74	66.28
8.18	139	69	73.97	12.48	141	76[11]	63.93
7.47	140	70	72.84	11.49	142	78	61.80
6.71	141	72	70.66	10.52	143	79	60.62
6.06	142	73	69.56	9.67	144	81	58.19
5.47	143	75	67.41	8.93	145	82	57.04
4.83	144	76	66.28	8.21	146	85	53.35
4.35	145	79	62.99	7.57	147	86	52.10
3.92	146	81	60.75	6.85	148	88	49.80
3.52	147	82	59.56	6.20	149	90	47.56

Some of the other models considered – for example model 6 where grade G was mapped to grade 1 – would exacerbate the situation further as the grade widths below grade 4 would be wider so the arithmetic position of grade 5 would be even higher, with a bigger difference in the standard between the two tiers at this grade.

These results may not be generalisable to other subjects, nor even to the new Mathematics GCSE, but the modelling does provide a stark warning that the grade 5 boundaries will not automatically be aligned. Therefore the grade 5 boundary on the foundation tier should be set using test equating both in the transition year and in subsequent years. Even then, as with untiered papers, there is no guarantee that the outcomes between awarding organisations will be aligned at what will be a key grade.

[11] The calculated mark for 2013 was 75.

The use of test equating to maintain standards between tiers is not without its detractors but it remains the most powerful tool we currently possess. To maximise its efficacy it is important that the test design is appropriate. There must be a sufficient overlap of marks between the two tiers, and although there is some debate over the amount of overlap needed, the general rule applied is that at least 20% of the marks on each paper should be common. It is also important that the items are targeted at the correct level of demand. Common items should be targeted at grades 4 or 5 as these are the overlap grades. There is a risk that, if awarding organisations are forced to raise the demand of their papers too far, the common questions may be inaccessible to candidates on the foundation tier, undermining the test equating and leading to spurious results. Therefore, whenever test equating is used, the boundaries should be confirmed using judgemental evidence, although such evidence is hindered by the Good and Cresswell effect, which postulates that, when making judgements, there tends to be a bias in favour of an easier challenge (Good and Cresswell, 1988).

Classification accuracy

Since the new grading scale has one extra grade, *ceteris paribus* the grade widths of boundaries will be narrower; so there is an increased risk of a candidate being awarded an incorrect grade. Since there are actually two grades being added above a grade C then the reduction in classification accuracy will be particularly acute at the higher end of the scale, although there could actually be improvements at the bottom end.

To see the possible effect that the new grading scale may have, it was applied to the higher tier of AQA's GCSE linear Mathematics from June 2013 (at subject level). Figure 6 shows the actual grade widths produced and the widths of the boundaries if the new grading scale had been used instead. The new grades are based on model 2c where grade C is mapped to grade 4, grade A is mapped to grade 7 and all other boundaries are set arithmetically with an allowed grade 3 being half the grade width of grade 4.

The new grading scale clearly adds extra discrimination as the A* candidates are split between grades 8 and 9 and the A candidates are split between grades 7 and 8. The grade 4 boundary is 21 marks wide whereas the grade C boundary is currently 31 marks wide.

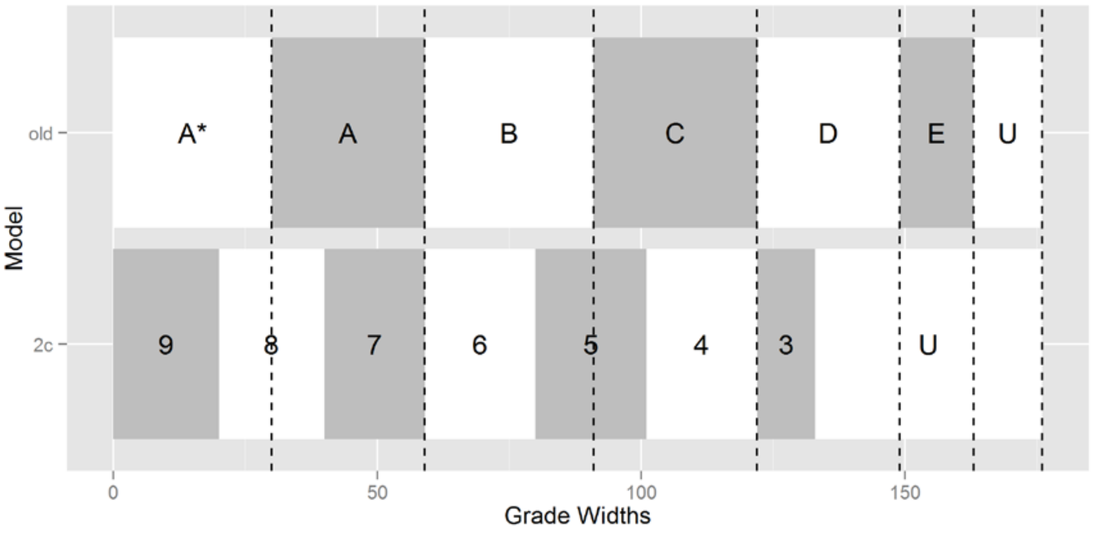


Figure 6 Grade widths for AQA's linear Mathematics GCSE (higher tier) in summer 2013

Classification accuracy was calculated using an item response theory (IRT) mixed format test approach. The model details are outlined by Lee (2008) but were fitted using the classify package (Wheadon, 2014) in R. For practical purposes it is based on a random sample of 500 candidates.

Table 4 shows the marginal classification accuracy (for the subject) under the two different grading scales with the associated standard error in brackets. The full results, including the classification accuracy by grade, are shown in Appendix 5. It can be seen that the probability of a candidate being awarded the correct grade falls from 0.822 to 0.743. Any candidate near a grade boundary is more likely to be misclassified, but it is reassuring to note that the probability of a candidate being misclassified by more than one grade remains very close to zero.

Table 4 Classification accuracy for AQA's linear GCSE Mathematics from summer 2013.

Grade Scale	Classification Accuracy
Current (A*-E)	0.822 (0.006)
New (9-3)	0.743 (0.005)

Classification accuracy is affected by numerous factors so any losses due to the increased number of grades could be offset by other changes in the design of the specification, the assessment and the marking. The new specifications have been designed to be more demanding so it is likely that the grade 4 boundary will be lower than the current C, leading to more marks between grades 4 and 7 than between the current C and A. AQA's new Mathematics GCSE has 240 marks, an increase of over 40%; this should ensure grade boundaries are kept a fair distance apart. So, in the case of Mathematics the changes to the assessment structure may have mitigated the negative impact on classification accuracy that would have arisen from having an extra grade.

However, subjects that will be untiered (particularly those which are tiered currently) are less likely to have more marks for the number of grades so may be more prone to increases in grade misclassifications. This is particularly true for English and English Literature, subjects which naturally tend to be less reliable than Mathematics. There is a risk that the grading reliability of these new qualifications could be severely compromised. Great efforts will need to be made when setting questions and mark schemes, and in the standardisation of markers, to ensure that the marking is as reliable as possible and that the full mark range is used so that boundaries can be spread out as much as possible.

Conclusions

This paper and the simulations presented have only looked at the first awards of the reformed GCSE specifications and do not deal with issues regarding subsequent awards. It is important to note that any grading scales will present difficulties, either narrow grade widths or sparsely populated grades, if the mark distribution fails to adequately spread out the candidates.

The new qualifications need to be carefully designed, as the advent of more grades in the new grading scale is likely to increase the risk of grade misclassifications; the risk is particularly acute at the higher grades. However, the first raft of new GCSEs have already been designed and it is possible that the new design may lead to compressed mark distributions and an increased risk of grade misclassifications.

By aligning the bottom of grade G with the bottom of grade 1, Ofqual has ensured that the new GCSEs will be accessible to the same percentage of candidates as currently. However, as there will be fewer grades at the bottom end of the range to discriminate between candidates, there is a chance that grade 3 will be an overly populous grade.

Statistical predictions and extrapolations are likely to be weaker at the extremes of the mark distribution (e.g. at grade 9) so there is a risk that standards may not be aligned between awarding organisations at these grades. Hopefully it will be possible to use pre-results statistical screening to ensure standards are equivalent across the awarding organisations at every grade. If this is not possible, then the use of predictions at every grade will need to be investigated.

Ofqual's decision to fix the percentage of grade 9s at 20% of the cumulative percentage at grade 7 ensures that it will not be a sparsely populated grade. However, it is possible that the grade boundaries will be narrower for the top grades so candidates could be misclassified more often. This approach to determining grade 9 also fails to account for differences in the candidature between subjects or between awarding organisations.

Ofqual has said that test equating will be used to set the boundaries for grades 4 and 5 for tiered subjects. It is necessary to use test equating for both boundaries as setting the grade 5 arithmetically is likely to lead to different standards between the tiers.

So far there has been no mention by the regulator of professional judgement in the awarding process of the first year of the new GCSEs, even in a confirmatory capacity. With the importance of ensuring that standards are maintained between tiers, it seems advisable to use professional judgement to augment the statistical evidence in the case of tiered specifications at the very least.

It is imperative that the awarding organisations, Ofqual and the DfE ensure that the new grading scale is fully explained to all stakeholders so that it can be interpreted appropriately. Otherwise it could unduly advantage or disadvantage candidates who have taken the reformed GCSEs and therefore have grades on the new scale.

References

Azzalini, A. (2014). The R 'sn' package: The skew-normal and skew-t distributions (version 1.1-0). URL http://azzalini.stat.unipd.it/SN

Department for Education. (2011). *PISA 2009 study: how big is the gap? - A comparison of pupil attainment in England with the top-performing countries*. Retrieved May 29, 2014, from https://www.gov.uk/government/publications/pisa-2009-study-how-big-is-the-gap-a-comparison-of-pupil-attainment-in-england-with-the-top-performing-countries

Eason, S. (2014). *GCSE pre-results statistical screening*. Manchester, UK: AQA Centre for Education Research and Policy.

Good, F. J., and Cresswell, M. J. (1988). *Grading the GCSE*. London: Secondary Examinations Council.

Gove, M. (2013). Letter from Michael Gove to Ofqual sent 6 February 2013. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/278308/sos_ofqual_letter_060213.pdf

JCQ. (2013). GCSE and Entry Level Certificate Results Summer 2013. Retrieved from http://www.jcq.org.uk/examination-results/gcses

Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory*. (No. 27) CASMA Research Report. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Ofqual. (2011). *GCSE, GCE, Principal Learning and Project Code of Practice*.

Ofqual. (2014a). *Consultation on Setting the Grade Standards of new GCSEs in England*. Retrieved May 29, 2014, from http://comment.ofqual.gov.uk/setting-the-grade-standards-of-new-gcses-april-2014/

Ofqual. (2014b). *Board paper for new GCSEs in 2017*. Retrieved September 12, 2014, from http://ofqual.gov.uk/news/setting-standards-new-gcses-2017/

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Stringer, N. (2014). *Setting Grade A* Boundaries on the New Linear Qualifications*. Manchester, UK: AQA Centre for Education Research and Policy.

Wheadon, C. (2014). Classification Accuracy and Consistency under Item Response Theory Models Using the Package classify. *Journal of Statistical Software*, *56*(10), 1–14. URL http://www.jstatsoft.org/v56/i10/

Appendix 1: The Simulations

Input variables

R was used to calculate grade boundaries and the cumulative percentage of candidates at each grade for the new GCSEs based on simulations using the skew normal distribution.

The script has the following input variables:

- mean mark, μ
- standard deviation, σ
- skew parameter, α
- cumulative percentage of candidates achieving grades A, C and F
- model for mapping grades.

The maximum mark was set to 100. The minimum mark was set at zero. The number of marks generated for each distribution was 10,000.

Cumulative percentages at each reference point

The cumulative percentages were based on subjects from the June 2013 (All UK Candidates) JCQ Provisional GCSE (Full Course) Results. Table 5 shows the subjects used with their associated cumulative percentages. Subjects were chosen to give a range of values at the different grades.

Table 5 Cumulative percentages from JCQ figures from summer 2013

Subject	Cumulative Percentage (%)		
	A	C	F
All subjects	21.3	68.1	96.8
English	14.2	63.6	98.1
English Literature	22.8	76.8	98.3
Geography	27.0	69.0	97.0
Religious Studies	30.9	72.4	95.6

Grade boundaries (A*-G)

The grade boundaries for A, C and F were calculated from the cumulative percentages specified in Table 5. The other grade boundaries were calculated arithmetically using the current methodology as outlined in Ofqual's Code of Practice (Ofqual, 2011).

Appendix 2: The skew normal distribution

Mark distributions were generated using the skew normal distribution (Azzalini, 2014). The skew normal distribution has three parameters:

- a location parameter, ξ
- a scale parameter, ω
- a skew parameter, α (-∞ < α < ∞)

The scale parameter is given by the equation:

$$\omega^2 = \frac{\sigma^2}{1 - \frac{2\delta^2}{\pi}}$$

where

$$\delta = \frac{\alpha}{\sqrt{1 + \alpha^2}}$$

and the location parameter is given by the equation:

$$\xi = \mu - \omega\delta\sqrt{\frac{2}{\pi}}$$

When α = 0 the distribution is a normal distribution.

The skew normal distribution is continuous and unbounded. When generating numbers from a skew normal distribution for the simulations they were rounded to the nearest integer. Any numbers above the maximum mark or below zero were deleted.

Appendix 3: Model Details

Extensive modelling was carried out to inform the equivalences between the old and new grading structures. Only a few of the models are discussed in detail in this report but for completeness, and in order to understand the model numbering, a full list of all simulated models is given below.

In all models grade C is mapped to grade 4.

Model 1:

- Grade A is mapped to grade 7
- Grade F is mapped to grade 2.

Model 2:

- Grade A is mapped to grade 7
- Grade F is mapped to grade 1.

Model 3:

- No other fixed points.

Model 4:

- Grade A is mapped to grade 6
- Grade F is mapped to grade 2.

Model 5:

- Grade 7 is four fifths of the distance from C to A
- Grade F is mapped to grade 1.

Model 6:

- Grade A is mapped to grade 7
- Grade G is mapped to grade 1.

For each of the six models there were several possible variants (A-E), not all of the 30 possible combinations of models were actually used for the simulations.

Model A:

- Intervening boundaries are calculated by interpolation with grades 8 and 9 interpolated between the grade 7 boundary mark and the maximum mark.

Model B:

- Intervening boundaries are calculated to have equal percentages.

Model C:

- Intervening boundaries (2, 3, 5 and 6) are calculated by interpolation
- Grades 8 and 9 are calculated arithmetically, i.e. grade 8 is the same number of marks above grade 7 as 7 is above grade 6, unless the width between the maximum mark and the grade 7 boundary is less than three times the width between grade 6 and 7, in which case grades 8 and 9 are interpolated between grade 7 and the maximum mark.

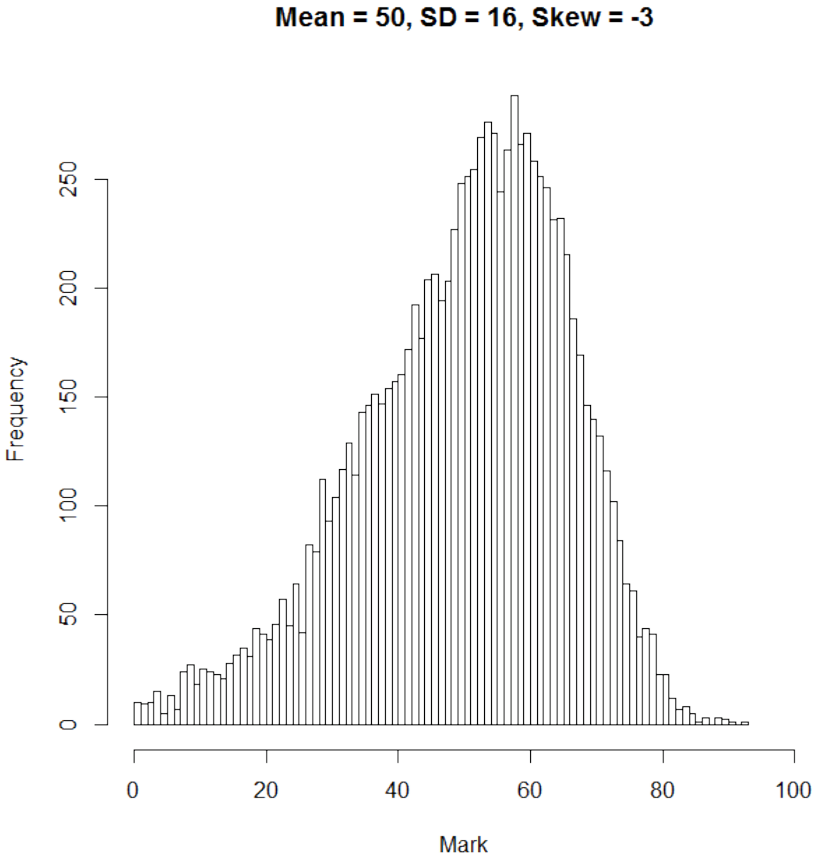
Model D:

- Intervening boundaries (2, 3, 5 and 6) are calculated by interpolation
- Cum % at grade 9 = 0.5 x cum % at grade A*
- Grade 8 is halfway between 7 and 9.

Model E:

- Intervening boundaries (2, 3, 5 and 6) are calculated by interpolation
- Cum % grade 9 = 0.2 x cum % grade 7
- Grade 8 is halfway between 7 and 9.

Appendix 4: Example of a mark distribution with negative skew



Appendix 5: Classification accuracy

Current Grades (A*-E)

- Marginal Classification Accuracy: 0.822 (0.006)
- Marginal Classification Consistency: 0.749 (0.007)
- Kappa: 0.661 (0.009)
- Marginal False Negative Error Rate: 0.092 (0.006)
- Marginal False Positive Error Rate: 0.086 (0.005)

Table 6 Accuracy by grade

Grade	Accuracy	False positive	False negative	Consistency
U	0.86	0.00	0.14	0.81
E	0.74	0.08	0.18	0.63
D	0.82	0.06	0.12	0.75
C	0.83	0.09	0.08	0.76
B	0.83	0.10	0.07	0.76
A	0.82	0.14	0.04	0.75
A*	0.79	0.21	0.00	0.77

New grades (9-3)

- Marginal Classification Accuracy: 0.743 (0.005)
- Marginal Classification Consistency: 0.647 (0.005)
- Kappa: 0.569 (0.006)
- Marginal False Negative Error Rate: 0.128 (0.006)
- Marginal False Positive Error Rate: 0.129 (0.007)

Table 7 Accuracy by grade

Grade	Accuracy	False positive	False negative	Consistency
U	0.89	0.00	0.11	0.84
3	0.56	0.21	0.23	0.45
4	0.75	0.12	0.13	0.64
5	0.74	0.15	0.12	0.63
6	0.74	0.15	0.11	0.64
7	0.73	0.17	0.10	0.62
8	0.77	0.17	0.06	0.69
9	0.71	0.29	0.00	0.69