CAMBRIDGE ASSESSMENT

Exploring the value of GCSE prediction matrices based upon attainment at Key Stage 2

Tom Benton and Tom Sutch

Cambridge Assessment Research Report

20th May 2013

University of Cambridge
Local Examinations Syndicate

Introduction

Prediction matrices are a key part of Ofqual's *Comparable Outcomes* approach to standard setting (http://www.ofqual.gov.uk/files/2012-05-09-maintaining-standards-in-summer-2012.pdf). The methodology consists of attempting to predict the likely grade distribution of a set of candidates given data about their prior attainment. This predicted (or *putative*) grade distribution is then used as one of the key pieces of information in deciding upon where grade boundaries should be set. If it is found (after setting grade boundaries) that the proposed grade distribution for candidates differs too far from the expected distribution given their prior attainment then grade boundaries may need to be adjusted accordingly.

The aim of this report is to provide further evidence regarding the effectiveness or otherwise of prediction matrices based on prior attainment at key stage 2 (KS2) for setting standards at GCSE. There has been some concern over using key stage 2 results as the basis to set standards for a number of reasons. In particular it has been suggested that due to the time gap between KS2 and GCSE, KS2 results are not a strong enough predictor of likely GCSE attainment. Furthermore, KS2 only covers three subjects (English, Maths and Science) and may not be a reliable predictor of performance in other subjects.

The specific aims of the research are:

- To evaluate the correlation between KS2 and GCSE results for all full GCSE subjects.
- To explore whether there are any obvious signs of KS2-based putative grade distributions being inappropriate in the presence of low correlation between KS2 and GCSE.
- To explore the standard errors of putative grade distributions as well as other evidence relating to the likely margin of error associated with the method.
- To investigate whether using additional information in addition to KS2 scores (specifically gender) improves the likely accuracy of the method.

Data

All of the data used for analysis comes from the National Pupil Database (NPD). For different analyses GCSE outcomes in years between 2008 and 2011 are considered. Analysis is restricted to the achievement of full GCSE qualifications by year 11 candidates. Any GCSEs achieved by this group of pupils prior to year 11 are excluded from analysis. Analysis is further restricted to candidates with a known final grade, that is, candidates under suspicion of malpractice or where their final GCSE grade is not provided for any other reason are also excluded from analysis. Data from all awarding bodies is included in the analysis.

For each GCSE taken by each candidate the following variables are defined (where available) for the purposes of analysis:

- Concurrent attainment measured by **Average GCSE grade** in subjects other than the one under investigation. For the purposes of calculating this variable, GCSE grades were converted into a numerical scale between 0 and 8 (with 8 being equivalent to A*). Only full GCSEs were included in the calculation of this measure. This measure was only calculated for candidates taking at least 4 full GCSEs[1].
- Prior attainment measured by **Average KS2 level** across each of English, Maths and Science. Sublevels were not used in the calculation of this measure. Key stage 2 levels below 2 were not included in this measure as these are not valid levels. Average KS2 levels were only calculated for candidates with valid levels across all three KS2 subjects.
- Prior attainment measured by **Average KS3 level**. Calculated as for KS2 but with levels ranging from 2 to 8. KS3 data was not available for any students taking GCSEs after 2010.
- Future attainment as measured by **AS level grade** in the same subject. This was only available for pupils who had continued to higher level study in the same subject and was

[1] That is, it is the average grade across at least 3 full GCSEs in addition to the GCSE with which it is concurrent.

Standard errors of putative grade distributions based upon KS2

This next section is concerned with estimating the standard errors of putative grade distributions derived from KS2. The method used to achieve this is that of balanced repeated replication (BRR) and closely follows the methodology applied by Benton and Lin (2011)[2]. This method estimates how much we might expect the predicted grade distribution to differ from the actual grade distribution given the sample sizes involved[3], if the assumptions underlying the technique are correct[4].

In calculating the standard errors it is important to remember that these come from two sources:

1. **Model standard errors**. These represent the uncertainty in predictions arising from the fact that the data analysed in one year only provides an estimate of the percentage we expect to achieve each grade given prior attainment. In other words, each cell of the prediction matrix has a standard error associated with it due to the fact it is based upon a finite sample of students. This source of error relates to uncertainty within the prediction matrices themselves and will be largely dependent upon the amount of data used to construct them.

2. **Innate standard errors**. Even if the expected percentage to achieve a given grade within each level of prior attainment is known precisely there is still uncertainty surrounding the numbers that will actually achieve this grade[5]. These standard errors will be largely dependent upon the number of students to whom each predicted grade distribution is applied.

The overall standard error of the putative grade distribution is found by combining these two elements.

This procedure was applied to find the standard errors for the KS2-based predicted grade distributions for all GCSE subjects taken in 2011. The data was restricted to those pupils with valid KS2 scores. In common with current practice in live awarding, candidates in either independent or selective schools were removed from analysis. The prediction matrices under consideration were constructed using data from candidates completing GCSEs in 2010. Estimates of innate standard errors were developed by applying balanced repeated replication to the achievement of candidates in 2011.

Once overall standard errors had been calculated it was possible to convert these into recommended tolerances for the method. This was done on the basis that tolerances would be based on 75% confidence intervals and thus could be generated by multiplying the estimated standard errors by 1.15. Table 3 shows the average recommended tolerance dependent upon the number of candidates entering any given subject with a particular awarding body.

Table 3 clearly shows that as the number of candidates entering a subject increases, the standard error (and hence the recommended tolerance) of the prediction matrices method falls. It can also be seen that the standard errors tend to be larger at grades C and A than at grade F. This is likely to be connected to the fact that the percentage of students expected to achieve

[2] See appendix 5 of Benton, T. and Lin, Y. (2011) Investigating the relationship between A level results and prior attainment at GCSE. Ofqual: Coventry.

[3] And imagining that KS2 predictions were not used to fix the subsequent actual distribution of grades.

[4] Namely that the relationship between key stage 2 attainment and grade achieved in any GCSE subject does not change between the year from which data is used to construct prediction matrices and the year in which they are applied.

[5] To illustrate the difference between an expected percentage and an actual percentage consider the example of tossing a fair coin. We know that the expected percentage of times that we will get heads is 50 per cent. However, random chance also plays its part and the actual percentage of times that we get heads may be somewhat different to 50 per cent (particularly for a small number of coin tosses).

Table 4: Regression of standard errors of predictive variables at each of grades F, C and A

	Grade F		Grade C		Grade A	
Independent variables	**Coefficient**	**Standard Error**	**Coefficient**	**Standard Error**	**Coefficient**	**Standard Error**
Intercept	0.05	0.03	0.22	0.08	0.21	0.05
V1	79.95	16.86	47.20	16.97	46.52	12.27
V2	161.76	12.10	197.79	10.88	175.03	8.12
Fit statistics						
R square		**0.80**		**0.81**		**0.88**
Correlation between predicted and original SEs		**0.89**		**0.90**		**0.94**

Further examining these results we can see that the number of pupils to whom the prediction matrix is being applied (as contained within V2) is a more influential factor on the likely standard error than the number of pupils used to construct the prediction matrix. Nonetheless, it can be seen that both of these factors are important in determining the likely standard error and therefore margins of error should not be based purely on the numbers of students to whom the matrix is being applied.

As a final piece of analysis, the correlations calculated in the previous section were added in to the models displayed in table 4 to ascertain whether there was a link with the margin of error of prediction matrices. Correlation between KS2 achievement and the grade achieved in a GCSE subject was not found to be a significant predictor of standard error. This may appear a surprising result. However, it is worth noting that standard errors are only concerned with the stability of a given estimate (that is, whether it would vary between samples). Provided we have a large sample, prediction matrices should provide stable estimates of grade distributions even if KS2 does not provide much in the way of useful information about future likely achievement. Furthermore, as shown in the previous section, the correlations between KS2 and GCSE grades are all within a fairly narrow range (largely between 0.5 and 0.7). There are no subjects where the kind of very large correlations occur that would allow the method to have very low standard error even in the presence of small sample sizes[7].

[7] Theoretically there should be a link between correlation and standard error in that if the correlation between KS2 and GCSE grade was equal to 1 then the standard error would be zero. In such an instance, regardless of which pupils are included in a sample, provided the distribution of prior attainment remained constant (which it does within the BRR method), the same predicted grade distribution would be generated. However, no subjects have a correlation with KS2 grades anywhere near 1 and so we are not able to detect such an effect within our data.

Alternative formulations for the margin of error of KS2 based putative distributions?

Using the calculations in the previous section as a basis for estimating the real margin of error of prediction matrices is predicated on the belief that the foundational assumptions of the method are correct. Another way to examine the margin of error in the prediction matrices methodology is to compare the results from this method to the results that would be gained if we were able to use a far more powerful variable in setting grade thresholds – namely, concurrent GCSE attainment. We have already seen that the average correlation between concurrent GCSE attainment and the grade achieved in any individual GCSE subject is much higher (at around 0.7) than the average correlation with KS2 (at around 0.5). For this reason it is reasonable to assume that if this information were available at the time of standard setting (which, of course, it cannot be) we would certainly prefer the use of this information to the use of KS2. Indeed predicted distributions based upon concurrent attainment are one of the key ways in which inter-board differences are ultimately evaluated post awarding. Thus we can evaluate the accuracy of KS2-based putative grades by comparing the predicted distributions produced by this method to those predicted by a similar method based upon concurrent GCSE.

Initially this was done as follows:

- Restrict the data to candidates with both concurrent GCSE and KS2 data available and exclude candidates in independent or selective schools.
- Use historical data from 2010 to estimate the predicted distribution of grades in 2011 in terms of the cumulative percentage of candidates to achieve F or above, C or above and A or above. This can be done using the relationship between GCSE grades in individual subjects and either prior attainment at KS2 or concurrent attainment in other GCSEs.
- Compare the two predicted distributions and evaluate the average size of absolute differences in putative grades and how this varies dependent upon sample sizes and correlation with KS2 as evaluated earlier.

For the purposes of the above analysis, concurrent attainment at GCSE was broken into 11 categories by rounding mean GCSE grade to the nearest 0.5 and combining all candidates with a mean GCSE score of 3 or less into a single category. This method was preferred to the method of splitting GCSE scores into deciles as it was found that it gave results that were a little closer (and in some instances substantially closer) to the actual grade distributions of the awarded grades in 2011. For simplicity, analysis was also done combining data across all awarding bodies rather than examining each one separately. The results are shown in figures 1 and 2.

Figure 1 shows the association between the log (base 10)[8] of the number of candidates entering a subject and the absolute difference in the predicted grade distributions generated by KS2 and concurrent GCSE attainment. In general it can be seen that as the sample size increases the difference between the two techniques decreases. Subjects with at least 1000 entries (3 on the x axis) tend to provide estimates within 2 percentage points of each other. Subjects with a least 10,000 entries (4 on the x axis) tend to provide estimates that are within 1 percentage point of one another. These results are broadly in line with the recommended tolerances presented in table 3.

However, this comforting picture is somewhat disturbed by the presence of a few highly noticeable outliers within the chart. The largest difference between predicted grade distributions is found at grade C for Bengali[9], although this is a very small entry subject with only 507 candidates included in analysis. Of more concern is apparent difference between distributions for the three single sciences; apparent as the outliers towards the right hand side of figure 1.

[8] This means that a value of 3 on the x axis indicates 1000 pupils entering a subject, 4 indicates 10,000 entering a subject and 5 indicates an entry of 1000000.

[9] The second largest difference is at grade A for Bengali.

Figure 1: Association between number of entrants to particular subjects and the differences between putative grade distributions generated from KS2 and concurrent GCSE results

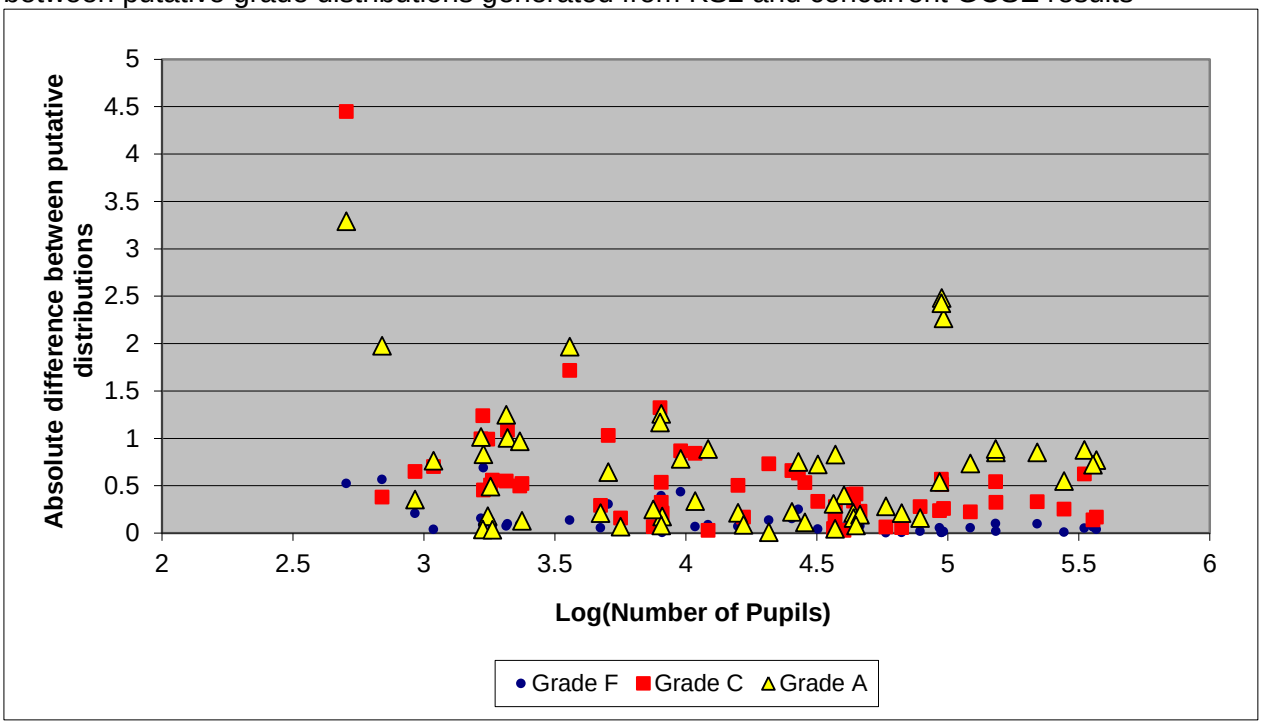
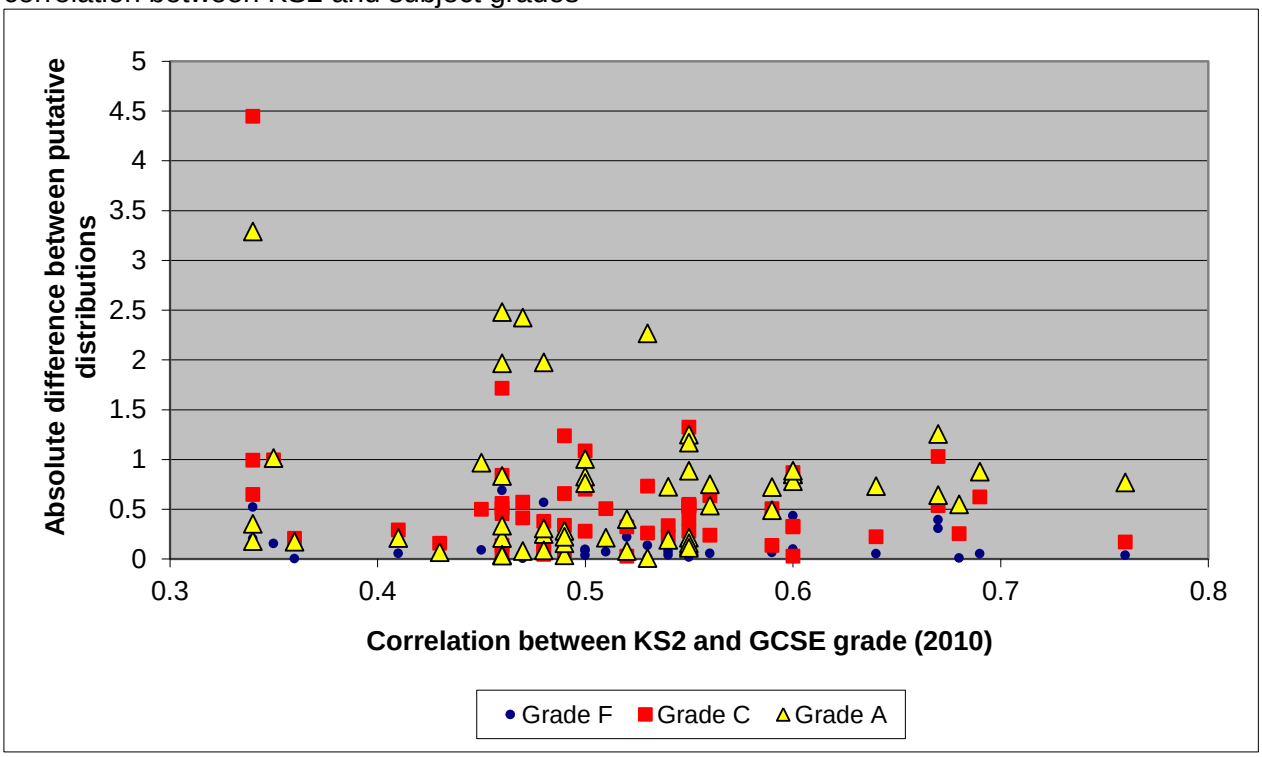


Figure 2 illustrates the association between the difference between putative grade distributions and the correlation between KS2 and subject grades. A slight relationship is apparent between the KS2 correlation and the method providing very different results to a method based upon concurrent GCSE results; however, on inspection this could be because the log of the number of candidates and the KS2 correlation are themselves correlated.

Fitting a linear regression of each of the differences at A, C and F on the log of number of candidates and KS2 correlation showed that, once the size of entry had been controlled for, the effect of the KS2 correlation was not significant. This indicates that whatever problems there may be at the heart of the use of KS2 to set GCSE standards, low correlation cannot be definitively identified as the source.

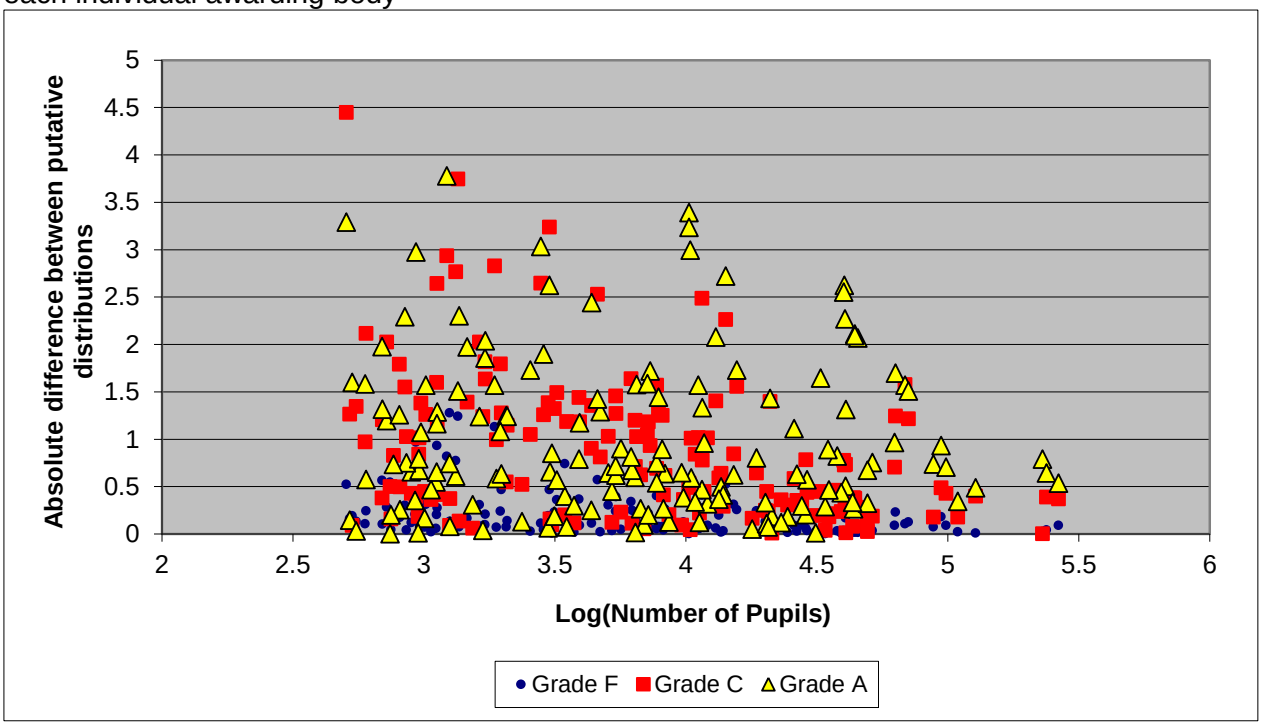
Figure 2: The association between difference between putative grade distributions and the correlation between KS2 and subject grades



The analysis in figure 1 and figure 2 considers the differences between predicted grade distributions across all boards. Figure 3 examines the differences between predicted grade distributions based upon KS2 and mean concurrent attainment within each individual board for each GCSE subject. Two things can be noted from this data. Firstly, whilst there is a general tendency for the differences between the methods to reduce as the sample size increases, the differences remain a little larger than might be expected from the tolerances displayed in table 3. Specifically we can see that for subjects with entries of more than 10,000 candidates the difference between the predicted grade distribution from KS2 is regularly more than 2 percentage points away from the grade distribution predicted by concurrent attainment. This indicates that whereas the assumptions of prediction matrices may hold relatively well when applied across boards (as shown by figure 1), they may apply less well when applied within individual boards. This suggests that current tolerances for the putative grade distribution are set too low; particularly for subjects with large numbers of entries.

The second thing that can be noted from this data is that whereas in figure 1 only a few outliers could be seen, a greater number of apparent outliers are evident when we look within individual boards. This implies that, whilst prediction matrices based on KS2 will generally provide reasonable results, they do not provide an infallible source of information. This will be explored further in a later section.

Figure 3: Association between number of entrants to particular subjects and the differences between putative grade distributions generated from KS2 and concurrent GCSE results within each individual awarding body



Is there any evidence of setting grade boundaries using KS2 results affecting the predictive validity of GCSEs?

A possible concern with using KS2 grades to set GCSE standards is that, if this method led to incorrect standards being set, a given level of achievement at GCSE would no longer have the same meaning in terms of future predicted attainment at AS level (or beyond). This would have serious implications for the validity of GCSE results as it would mean that decisions about the likelihood of candidates being able to cope with the demands of higher level qualifications would be wrongly assessed and decisions about candidates' suitability for further study may be incorrect.

In order to assess any issues with predictive validity the following method was used:

- Restrict the data to candidates with KS2 data available[10] that also went on to study a subject that they took at GCSE at AS level.
- Using historical data from 2008 GCSE achievement (and subsequent AS level achievement in 2009 or 2010) to estimate the predicted distribution of grades in 2009 in terms of the cumulative percentage of candidates to achieve F or above, C or above and A or above. This can be done using the relationship between GCSE grades in individual subjects and either prior attainment at KS2 or future grade in the same subject at AS level.
- Compare the two predicted distributions and evaluate the average size of absolute differences in putative grades and how this varies dependent upon sample sizes and correlation with KS2 as evaluated earlier.

Note that the predictive validity of GCSEs (that is, predicting forwards in time to AS level) can be evaluated by the relationship going in the other direction (that is, going back in time from AS level to GCSE). For example, if GCSEs were to become easier and the same level of GCSE attainment became associated with lower attainment at AS level, we would notice this in the reversed relationship as the same AS level grades would become associated with higher average GCSE grades.

Note that the sample sizes available for this analysis are much smaller than for any of the analyses described previously; only a minority of pupils will go on to study the same subject at AS level. Furthermore, since only the highest performing pupils will pursue further study in the same subject at AS level, almost all candidates within the data used for this analysis achieved at level C or above at GCSE. For this reason analysis is restricted to differences in the predicted percentage to achieve A or A* at GCSE on the basis of either prior attainment at KS2 or future attainment at AS level. This means that only 30 subjects are included within this analysis. Results are shown in figures 4 and 5.

Given the small number of subjects available for analysis, firm conclusions are tricky. Nonetheless, in both cases there does appear to be a slight association. Figure 4 shows that as the sample available for such analysis increases the differences between the different putative distributions decrease. Furthermore, the size of these differences is broadly in line with the recommended tolerances estimated earlier in table 3.

Figure 5 also shows a hint of a relationship where those subjects with a larger correlation with KS2 display smaller differences between putative grades based on AS results. There is a single large outlier towards the right hand side of the graph[11]. This relates to General Studies – a somewhat unusual AS level subject. Without this subject the relationship would be revealed even more clearly.

[10] Note that, for the version of the NPD used in this analysis, pupils from independent schools were not available. Data from pupils within maintained selective schools was available and was included within analysis in order to allow the greatest possible amount of data to be included.

[11] Correlation with KS2 of 0.68 in 2009, difference in grade distributions of 2.4 percentage points.

Figure 4: Association between sample size and the differences between putative grade distributions generated from KS2 and AS results

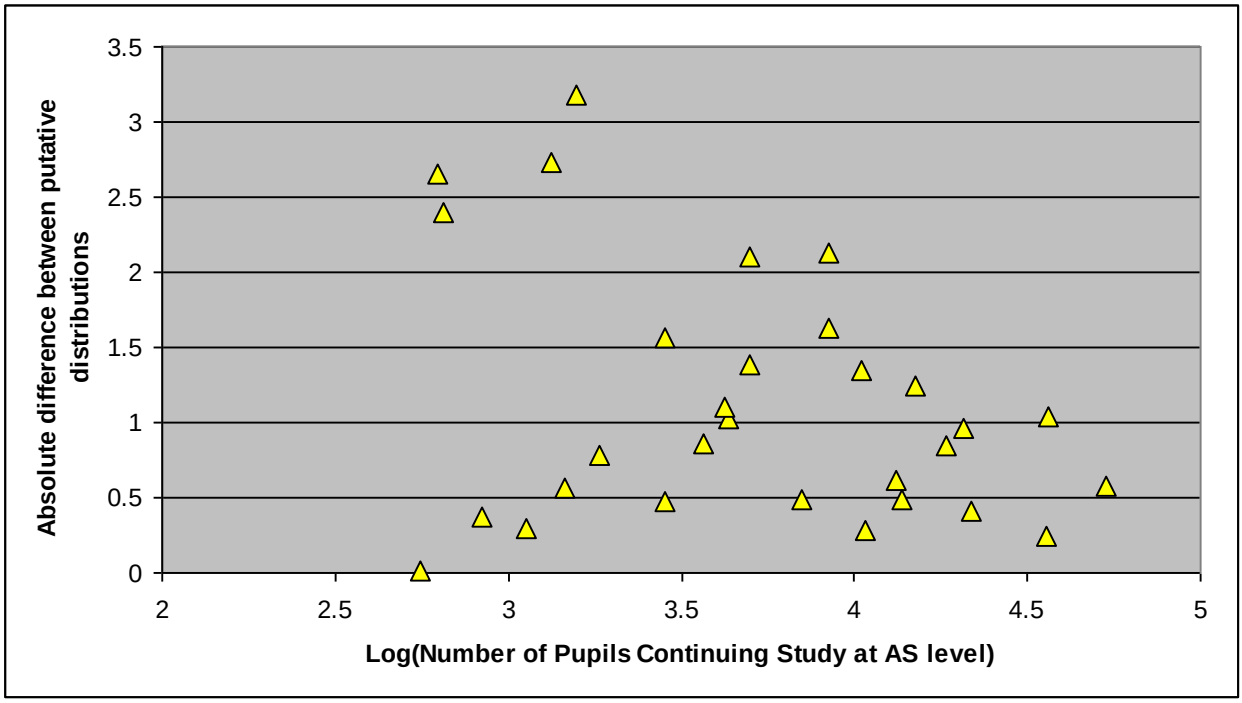
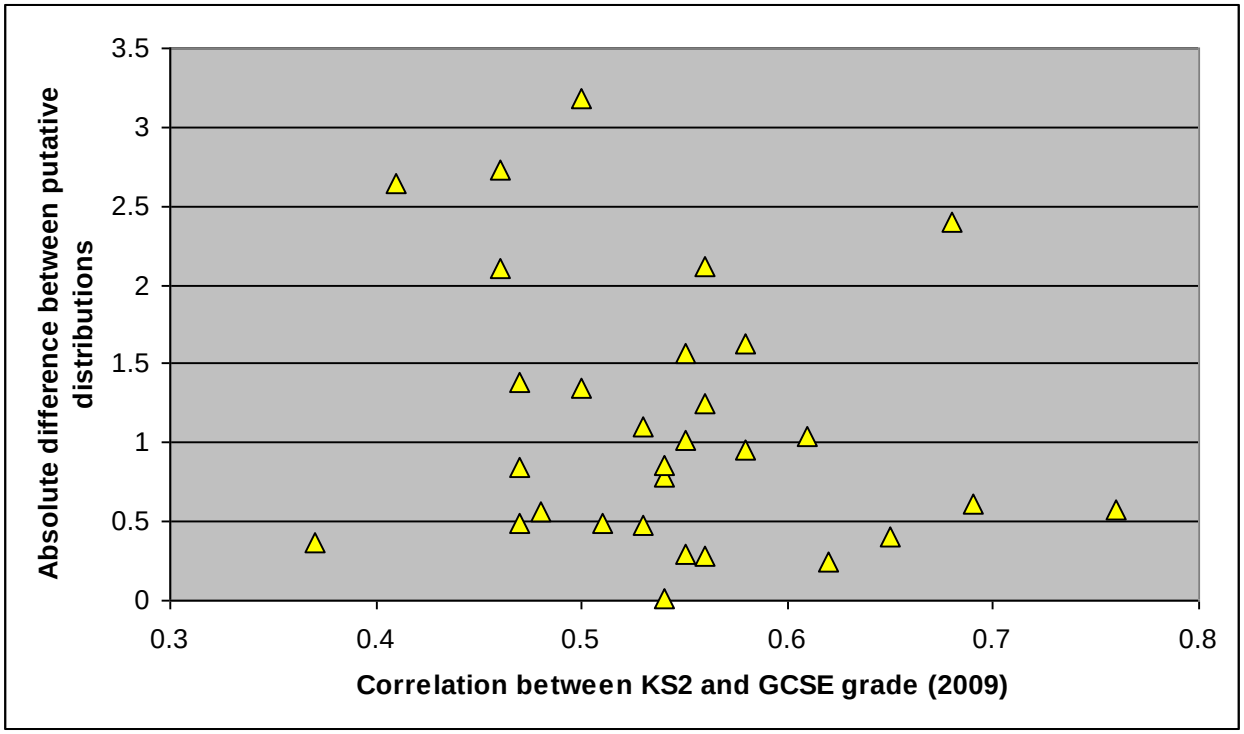


Figure 5: The association between KS2-GCSE correlation and the differences between putative grade distributions generated from KS2 and AS results



Having said the above, neither of the relationships visible in figures 4 and 5 are statistically significant. The number of subjects studied is too small to be certain that the pattern revealed above is not simply the result of random fluctuation. Nonetheless, the results in this section may provide some interesting avenues for thought and further study.

What causes standards set using KS2 results to lead to large inter-board differences?

In practice it has been noticed that the introduction of increased usage of prediction matrices in standard setting has not led to major improvements in inter board comparability when this is later assessed using concurrent attainment.

In order to fully study this issue it is important that data is not simply restricted to those candidates with matching KS2 information. Standards set using putative grades from matched data are subsequently applied to the entire GCSE cohort regardless of whether they have matching KS2 data; that is, the same grade boundaries are used for both sets of pupils. For this reason a slightly different approach to previous sections was adopted to allow us to identify subjects where the use of prediction matrices would lead to very different results to those that might be suggested by the use of a more appropriate data source (if it were available); namely, concurrent GCSE attainment. The procedure for analysis was as follows:

- Restrict data to candidates with matching concurrent GCSE attainment; that is, candidates that have taken at least 3 other GCSEs beyond the GCSE subject being studied.
- Restrict 2011 data to OCR candidates and match in data about the UMS score of candidates.
- Restrict analysis to the 52 GCSE subjects with at least 500 year 11 candidates taking the subject with OCR in 2011.
- Generate putative grade distributions for 2011 OCR candidates using historical data from 2010[12] and based on four different possible data sources:
    - Mean concurrent GCSE. As with the analysis in an earlier section this was split into 11 categories prior to analysis.
    - Key stage 2. Note that since not all pupils have matched KS2 data[13] this is a two stage procedure. First, the putative percentage is calculated for matched candidates in non-selective and non-independent schools. Next, grade boundaries on the UMS scale[14] are identified that would yield these putative grades. Finally, these grade boundaries are applied across all pupils[15] (matched and unmatched) to yield an overall putative grade distribution.
    - Common centres. That is, results within 2011 OCR centres in the same GCSE subject in 2010 (regardless of which board they were with in 2010). For each OCR centre, the probability of 2011 pupils achieving any grade was estimated to be equal to the percentage of pupils in the centre who achieved that grade in 2010. Since a small number of centres will not have historical information a similar two-stage procedure was used as for key stage 2.
    - Reproducing the cumulative percentage for OCR candidates in 2010 for 2011. That is, if 46.7 per cent of OCR year 11 Biology candidates were awarded a grade A/A* in 2010 then the putative percentage for 2011 will be exactly that (46.7).
- Compare the putative percentages from mean GCSE score to the putative percentages from the other three methods to provide an idea of the relative accuracy of each method.

The above methodology allowed us to perform two tasks. Firstly, by taking the predicted percentage from mean concurrent GCSE as a “gold standard”, the results of this analysis allowed us to compare the relative accuracy of the other three methods that could be used to set standards. Secondly, this analysis allowed us to identify any subjects where the standard implied by KS2 was out of line with the standard implied by other methods. We were then able to

[12] Across all boards

[13] And that, as within current practice in the use of KS2 prediction matrices, we exclude the KS2 results of candidates in independent and selective schools.

[14] Obviously in practice we cannot directly manipulate grade boundaries on the UMS scale. However, for the purposes of a research project this would seem like a reasonable procedure.

[15] Including those in independent and selective schools.

explore whether there were any common patterns relating to big differences between the different methods.

An overall summary of the differences between the different methods is shown in table 4. The results suggest that on average the predicted grade distribution based on KS2 is closer to the predictions from concurrent attainment than either a common centres approach or simply carrying forward the percentage achieving particular grades from the previous year. Having said this, the improvement in accuracy, either in terms of mean or median difference across subjects, tends to be within 1 percentage point of the accuracy of the common centres approach. For example, at grade A, the median difference between KS2-based predictions and the gold standard is 1.0 percentage points compared to a median difference of 1.6 percentage points for the common centres method.

It is also worth noting that the common centres approach itself appears to provide a slightly more accurate method for settings standards than simply reproducing the pass rate from the previous year.

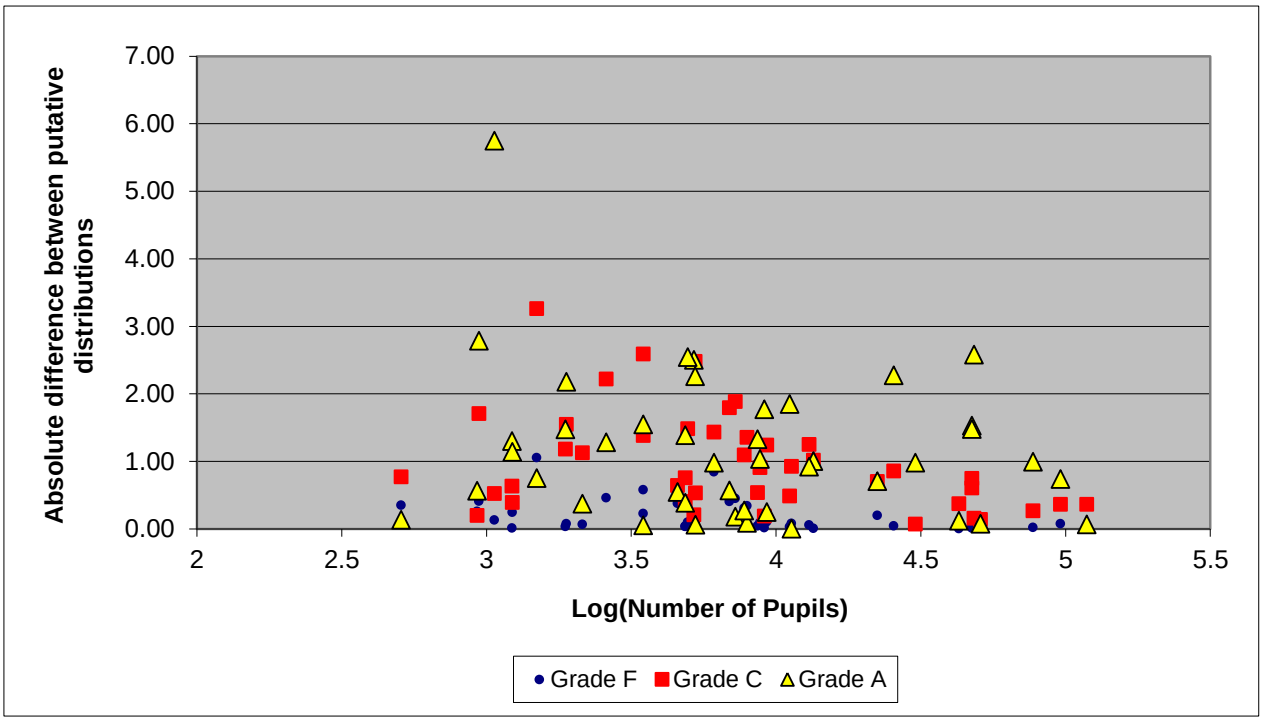
Table 4: Extent to which different methods match the predicted grade distributions generated using concurrent GCSE attainment

	Grade F			**Grade C**			**Grade A**		
	Absolute difference between putative percentage from mean GCSE and…			**Absolute difference between putative percentage from mean GCSE and…**			**Absolute difference between putative percentage from mean GCSE and…**		
Results across all 52 subjects	**KS2**	**Common Centres**	**Repeat 2010 results**	**KS2**	**Common Centres**	**Repeat 2010 results**	**KS2**	**Common Centres**	**Repeat 2010 results**
Mean	0.3	0.6	0.7	1.0	2.0	2.7	1.2	2.3	2.6
Median	0.1	0.3	0.5	0.8	1.2	1.8	1.0	1.6	1.8
Min	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.1
Max	2.9	3.8	2.5	3.3	10.7	9.8	5.7	13.4	10.2
Standard Deviation	0.5	0.7	0.6	0.7	2.2	2.4	1.1	2.3	2.5

The association between differences for individual subjects and the number of OCR entries in 2011 is shown in figure 6. Again we see a general tendency for the differences between the methods to reduce as the sample size increases. However, the differences between methods remain a little larger than might be expected from the tolerances displayed in table 3; particularly for subjects with large entries. Specifically we can see that for subjects with entries of more than 10,000 candidates the difference between the predicted grade distribution from KS2 is regularly more than the recommended tolerance of 1 percentage point away from the grade distribution predicted by concurrent attainment.

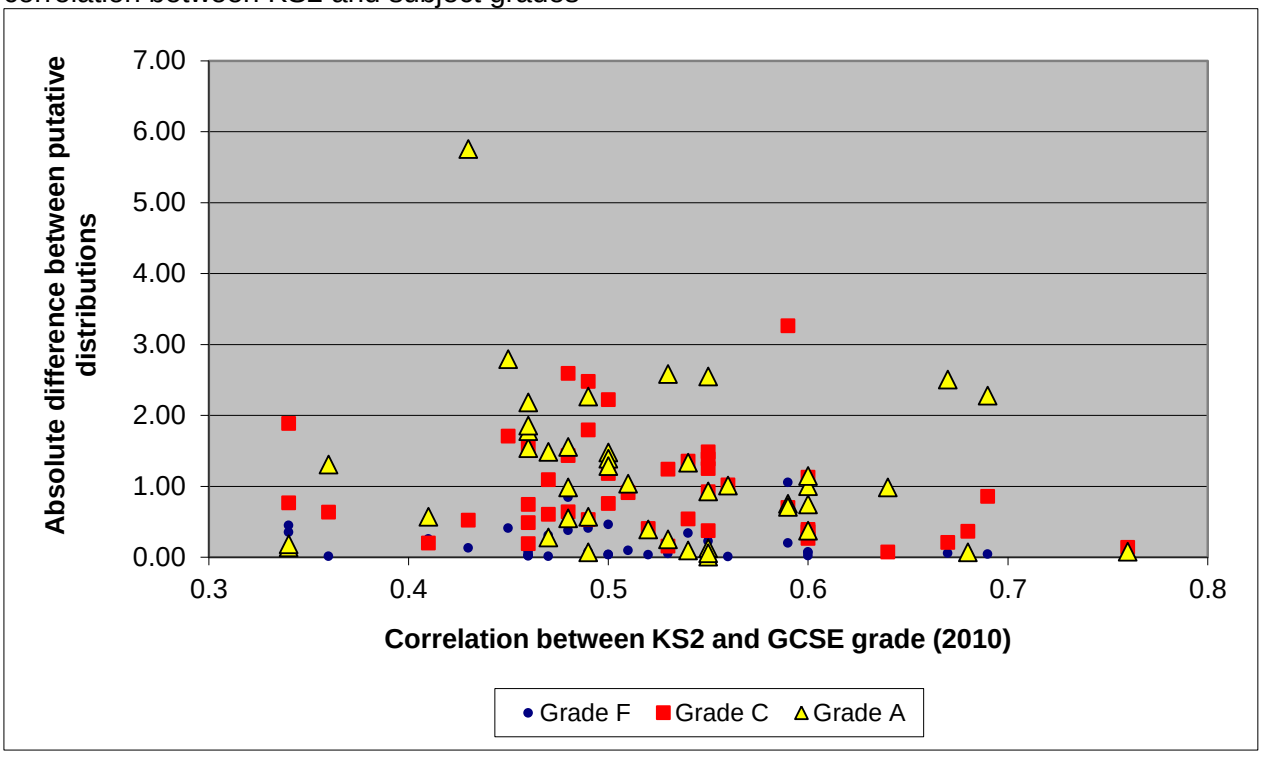
More pertinently it is worth noting that the current tolerance recommended by Ofqual for subjects with more than 3,000 entries is just 1 percentage point. Three thousand entries relates to a value of roughly 3.5 on the x-axis of figure 6. It can be seen that with this size of candidature, a difference between putative distributions exceeding this recommended tolerance is very nearly the norm rather than the exception.

Figure 6: Association between number of 2011 OCR entrants to particular subjects and the differences between putative grade distributions generated from KS2 and concurrent GCSE results



The association between differences for individual subjects and the 2010 correlation between subject grades and KS2 is shown in figure 7. Although there is a hint of an association in this chart, the sample size is too small to say with any certainty whether low correlation affects the accuracy of prediction matrices.

Figure 7: The association between difference between putative grade distributions and the correlation between KS2 and subject grades



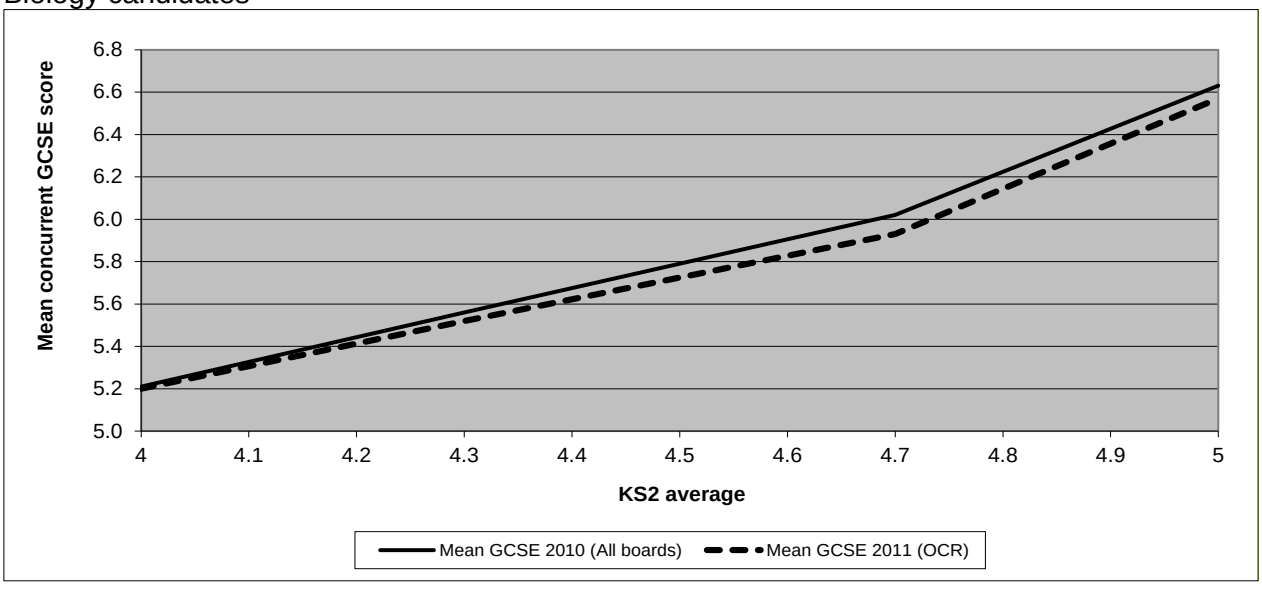
It is clear from the above analysis that any potential weakness in the validity of KS2-based predictions does not appear to be purely restricted to subjects with a low KS2-GCSE correlation. For this reason, more detailed investigations of the causes of the differences between predicted distributions based on key stage 2 and on concurrent GCSE are undertaken below.

The most obvious reason why the predicted distribution from key stage 2 may not match the predictive distribution from concurrent GCSE is that the two measures may be in line with one another. For example, it might be that KS2 suggests that AQA candidates are less able than OCR candidates whereas concurrent GCSE suggests the reverse. For our own analysis, what is important is whether the relationship between KS2 and concurrent GCSE remains constant between years. For example, if within 2011 the concurrent attainment of candidates is lower than we would expect given their KS2 achievement (and the kind of relationship between the two we have seen in the past), then KS2 achievement will tend to overestimate the likely achievement of the candidates. An instance of this is illustrated from Biology GCSE. Analysis of this OCR subject is based upon 48,334 entries in 2011 with a KS2 match rate of 94 per cent. The predicted percentage to achieve A or above from KS2 attainment was 46.3 per cent compared to 43.7 per cent based on concurrent GCSE achievement yielding a difference of 2.6 percentage points.

The reason behind the differences in the predicted grade distributions is shown in figure 9. Figure 9 shows the relationship between KS2 and mean concurrent GCSE (i.e. not in Biology) for Biology candidates in 2010 and 2011. For the purposes of this study the 2010 data is drawn from all boards (as would be the case in the application of prediction matrices in practice) and the 2011 data is based upon OCR data only. As can be seen, across the range of prior attainment, OCR Biology candidates in 2011 are marginally less able than candidates with similar prior attainment across all boards in 2010. That is, looking at their attainment in other GCSEs, they perform less well than would be expected. This implies that the use of KS2 prior attainment to set standards in this GCSE overestimates the true ability of OCR's candidates and will result in setting grade boundaries that are too lenient; in this case allowing an additional 2 per cent of pupils to achieve a grade A.

The differences in the concurrent attainment of candidates given their prior attainment are in fact very small; less than 0.1 grades. It is interesting to note that such small differences can lead to predictions based on KS2 being out of tolerance with predictions from concurrent attainment. This shows just how strongly the assumption of "all things being equal" needs to hold between one year and the next in order for results to be within the tolerances recommended by Ofqual.

Figure 9: The association between KS2 achievement and mean concurrent GCSE attainment for Biology candidates



A second possibility that may damage the validity of KS2-based predictions is that they require us to extrapolate results from one population (that is, candidates with matched KS2 information) to the population as a whole. An example of where this issue occurs is for Art & Design (Textiles). Analysis of this OCR subject is based upon 1,067 entries in 2011 with a KS2 match rate of 86 per cent. The predicted percentage to achieve A or above from KS2 attainment was 37.9 per cent compared to 32.1 per cent based on concurrent GCSE achievement yielding a fairly large difference of 5.8 percentage points.

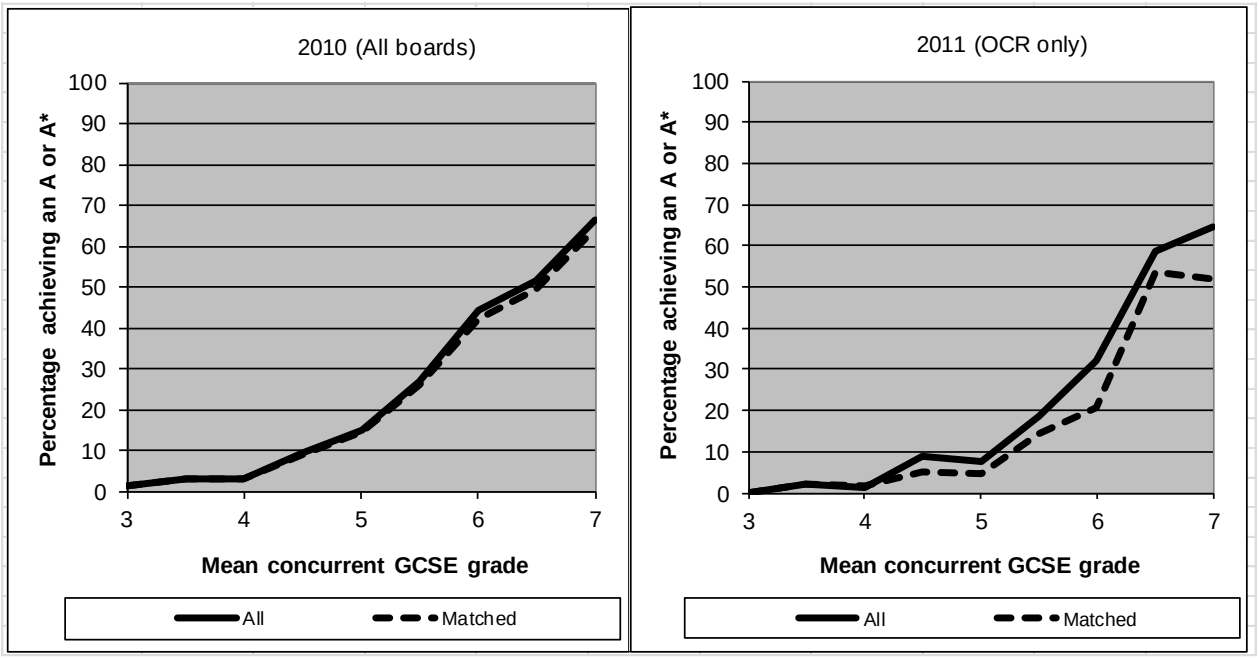
Firstly note that this difference is not due to the same mechanism shown in figure 9 (that is, OCR candidates in 2011 having apparently lower ability than those with similar prior attainment in 2010). The true reason for the difference is shown in figure 10. Figure 10 compares the association between concurrent GCSE grade and the chances of achieving grade A or above in GCSE Art & Design (Textiles). This relationship is shown both for the whole sample of pupils as well as restricted to those with matching KS2 achievement. This relationship is shown both in 2010 (using data from all boards) and in 2011 (using data from OCR only). The figure shows that, whereas in 2010 pupils with matched KS2 data tend to have similar performance compared to all candidates, in 2011 they underperform. This means that relying on the data from matched KS2 candidates in 2011 implicitly assumes that we expect them to be as strong as candidates in 2010. Given that matched candidates are now weaker than the unmatched candidates, this is unlikely to be the case in 2011 and so setting standards based on those with matching KS2 data will result in setting grade boundaries that are too lenient.

Another way to explain why differences between matched candidates and the cohort as a whole will lead to differences with the predicted distribution based on concurrent attainment is as follows:

- When we make a predicted distribution using concurrent attainment we use the whole cohort. This means we assume that pupils in 2011 are equally good at Art & Design (Textiles) relative to their other subjects as they are in 2010.
- When we make a predicted distribution using KS2 we use only the matched cohort. We are now assuming that *matched* pupils in 2011 are equally good at Art & Design (Textiles) relative to their KS2 (and by implication relative to their concurrent attainment) as they are in 2010.
- However, figure 10 shows us that the above assumptions cannot *both* be true as in 2010 there is no difference between the matched cohort and the cohort as a whole, whereas in 2011 there is.

Putting this in plain language, figure 10 shows that, relative to the cohort as a whole, matched candidates are not as good at Art & Design in 2011 as they were in 2010. However, our model assumes that they are just as good and so we would set grade boundaries that are too generous.

Figure 10: The association between mean concurrent GCSE grade and probability of achieving grade A or above in GCSE Art & Design (Textiles) for all and matched candidates in 2010 and 2011



Is there any value in taking account of gender as well as key stage 2 attainment?

One suggested way to improve the effectiveness of prediction matrices is to allow them to take account of variables beyond prior attainment at key stage 2. For example, it may be suggested that the prediction models could be improved by taking account of changes in the gender distribution of cohort as well as prior attainment at KS2. In order to test this idea, the following procedure was used:

- Restrict data set to candidates with information on their prior attainment at key stage 2 and gender.
- Fit two predictive models based on historical data in 2010 to produce the predicted probability of students having each grade in each subject in 2011. The first model should be based on KS2 prior attainment only (as for prediction matrices currently). The second method should be upon an ordinal logistic regression taking account of both KS2 attainment and gender. Compare the predicted probabilities to the actual outcomes of individuals to work out the deviance of each method.
- Compare the overall deviance for each method.

The overall results are shown in table 5. As can be seen the improvement in model fit from including gender is very slight indicating that this will lead to little practical improvement in the model.

Table 5: Overall deviance of models including and excluding gender across all subjects

Model	**Total deviance summed across all subjects**
Predictions based on KS2 only	188554
Predictions based on KS2 and gender	188255
Percentage improvement	**0.16%**

Additional analysis was also undertaken comparing the putative grade distributions from models including gender and not including gender for each GCSE subject within every awarding body. These comparisons found that, once analysis was restricted to subjects with at least 500 entries within a given awarding body, only one instance (Edexcel Urdu C grade) was found where there was a change of more than 1 per cent in the putative percentage. This finding confirms that taking account of gender would make little difference to the practical application of prediction matrices.

Summary

- GCSE subject grades have a lower correlation with KS2 than with either KS3 or concurrent attainment in other GCSE subjects. However, aside from subjects with very small entries, these correlations are rarely below 0.4. Equally they are rarely greater than 0.7 indicating that KS2 is neither a terrible nor an excellent predictor of likely achievement in any GCSE.
- Margins of error as calculated using balanced repeated replication indicate that current tolerances required by Ofqual may be marginally too low. Furthermore, comparing predicted overall grade distributions from KS2 with those based on concurrent attainment confirms that generally speaking differences between predictions from the two sources will be within expected limits given the standard errors. Using either method, the correlation between GCSE subject grades and KS2 does not appear to be a major factor in determining the margin of error.
- Having said this, greater discrepancies between the predicted grade distributions emerge once we look within individual awarding bodies rather than at overall achievement. This may indicate that the assumptions underpinning prediction matrices are less valid when applied within individual awarding bodies rather than across the whole population. It may further imply that the current tolerances applied to the technique are too low; especially for subjects with large entries.
- Although for the majority of GCSE subjects there are not huge problems with the use of KS2-based predictions[16], in a minority of cases these can yield different results to predictions based on concurrent GCSE attainment. This suggests that, whilst prediction matrices based on KS2 will generally provide reasonable results, they do not provide an infallible source of information.
- Examining the relationship between GCSE grades and future attainment at AS level reveals no evidence that setting grade boundaries using KS2 results affects the predictive validity of GCSEs. Furthermore, there is no evidence that such problems are particularly more likely to occur for GCSE subjects with lower correlation with KS2.
- On average, use of KS2 appears to provide more accurate estimates than either common centres or simply carrying forward the percentage from the previous year.
- The biggest differences between predicted grade distributions based on KS2 data and those based on concurrent GCSE results tend to relate to one of two causes:
    - Candidates with a particular board having higher concurrent attainment than would be expected given their prior attainment and the historical relationship between prior and concurrent attainment. This may occur if candidates with a particular exam board happen to attend schools with greater value added than others and can affect the accuracy of KS2 predictions matrices even if this effect is quite small.
    - The performance of candidates with matching KS2 data given their concurrent GCSE score and relative to the cohort as a whole changing between years. This cause may have particularly important implications when it comes to setting standards in summer 2015 as many more GCSE candidates will not have matching KS2 information due to the 2010 KS2 boycott. This may lead to change in the nature of the matched candidature relative to the unmatched group.
- Using information about candidate's gender in addition to their prior attainment appears to make very little difference to the predictions from the method.

[16] Although within individual awarding bodies differences may be a little outside the allowed tolerances.

Appendix 1 – Correlations between GCSE grade and other attainment measures by subject

	Correlations with KS2, KS3 and mean concurrent GCSE (excluding given subject)										
	2009				2010				2011		
Subject	KS2	Mean GCSE	KS3	N	KS2	Mean GCSE	KS3	N	KS2	Mean GCSE	N
Biology	0.58	0.84	0.77	67067	0.53	0.83	0.75	94463	0.51	0.83	121371
Chemistry	0.47	0.82	0.72	63274	0.46	0.82	0.71	91479	0.45	0.82	119119
Physics	0.47	0.8	0.72	63170	0.47	0.8	0.72	91442	0.47	0.81	118466
Science (Core)	0.71	0.84	0.85	365318	0.68	0.84	0.83	183096	0.69	0.83	303300
Science SA	-	-	-	-	0.71	0.83	0.86	154738	-	-	-
Additional Science	0.62	0.85	0.79	284239	0.6	0.84	0.78	259067	0.58	0.83	234685
Astronomy	0.53	0.78	0.71	685	0.48	0.74	0.66	837	0.46	0.75	881
Electronics	0.64	0.82	0.76	391	0.54	0.69	0.65	296	0.61	0.82	423
Environmental Science	0.59	0.78	0.72	2240	0.59	0.8	0.73	2036	0.68	0.81	1948
Geology	0.57	0.78	0.7	474	0.65	0.85	0.79	628	0.52	0.79	687
Mathematics	0.76	0.83	0.88	481464	0.76	0.83	0.88	453648	0.76	0.83	434018
Additional Mathematics	0.67	0.78	0.81	13087	0.67	0.76	0.8	10964	0.67	0.77	8817
Statistics	0.6	0.8	0.76	33302	0.56	0.79	0.74	32308	0.53	0.77	29788
Information & Communications Technology	0.53	0.74	0.65	44357	0.53	0.75	0.66	34333	0.52	0.74	26498
Motor Vehicle Studies	0.45	0.65	0.59	108	0.48	0.65	0.62	155	-	-	-
Business Studies:Single	0.56	0.8	0.71	62274	0.55	0.8	0.7	58680	0.6	0.83	48730
Business Studies & Economics	0.59	0.82	0.75	2017	0.55	0.81	0.73	1749	0.6	0.86	2968
Home Economics: Child Development	0.51	0.76	0.63	20822	0.51	0.75	0.65	19046	0.54	0.77	16475
Home Economics: Food	0.53	0.75	0.64	5986	0.55	0.76	0.66	6559	0.6	0.81	9237
Home Economics: Textiles	0.55	0.65	0.59	56	0.48	0.75	0.57	100	0.37	0.62	183
Art & Design	0.46	0.66	0.57	87824	0.46	0.66	0.57	83717	0.47	0.66	75056
Art & Design (Graphics)	0.41	0.61	0.5	4604	0.41	0.61	0.51	4643	0.42	0.61	5094
Art & Design (Photography)	0.37	0.59	0.46	5374	0.36	0.58	0.45	7185	0.38	0.59	8721
Art & Design (Textiles)	0.41	0.62	0.52	5701	0.43	0.63	0.53	6013	0.44	0.64	6358
Art & Design (3d Studies)	0.37	0.59	0.47	2288	0.34	0.57	0.45	1804	0.42	0.63	1980
Art & Design (Fine Art)	0.47	0.66	0.57	37015	0.46	0.66	0.57	36611	0.49	0.68	43960
Geography	0.65	0.86	0.79	136912	0.64	0.86	0.78	137457	0.67	0.87	142801
History	0.62	0.84	0.75	162033	0.6	0.84	0.75	164749	0.63	0.84	175775
Economics	0.54	0.8	0.71	1903	0.5	0.77	0.69	2131	0.55	0.82	2896
Humanities: Single	0.61	0.84	0.74	10833	0.6	0.83	0.73	8828	0.6	0.82	9948
Religious Studies	0.56	0.79	0.69	136284	0.55	0.79	0.69	139343	0.58	0.8	169845
Law	0.5	0.74	0.64	1490	0.46	0.7	0.6	1361	0.54	0.78	1761
Psychology	0.46	0.74	0.61	5203	0.48	0.76	0.64	5668	0.55	0.82	8175
Sociology	0.55	0.8	0.69	15135	0.48	0.78	0.65	15670	0.56	0.81	17014

	Correlations with KS2, KS3 and mean concurrent GCSE (excluding given subject)										
	2009				2010				2011		
Subject	KS2	Mean GCSE	KS3	N	KS2	Mean GCSE	KS3	N	KS2	Mean GCSE	N
English Language & Literature	0.69	0.86	0.8	440056	0.69	0.85	0.8	416852	0.71	0.85	400249
English Studies	0.51	0.68	0.67	2084	0.49	0.65	0.61	2612	0.51	0.63	2733
English Literature	0.61	0.8	0.72	402914	0.59	0.79	0.71	396660	0.59	0.79	402615
Drama & Theatre Studies	0.5	0.68	0.58	71789	0.49	0.67	0.58	67890	0.51	0.67	66137
Expressive Arts & Performance Studies	0.5	0.67	0.59	3822	0.45	0.66	0.55	3290	0.49	0.69	2545
Media/Film/Tv Studies	0.51	0.75	0.63	52062	0.49	0.74	0.61	50244	0.51	0.75	45343
Film Studies	0.68	0.85	0.77	391	0.55	0.79	0.67	1549	0.55	0.75	2498
Dutch	-	-	-	34	-	-	-	34	0.41	0.53	53
French	0.58	0.78	0.71	119660	0.56	0.78	0.7	114447	0.54	0.77	112583
German	0.54	0.76	0.69	54104	0.52	0.75	0.68	51977	0.52	0.76	49589
Italian	0.48	0.68	0.61	2090	0.46	0.69	0.58	2181	0.46	0.67	2184
Modern Greek	-	-	-	49	0.19	0.35	0.26	61	0.32	0.41	74
Portuguese	0.23	0.48	0.36	216	0.28	0.5	0.4	244	0.34	0.45	216
Spanish	0.53	0.74	0.67	39913	0.5	0.74	0.65	40820	0.5	0.74	47474
Arabic	0.09	0.21	0.17	438	0.05	0.18	0.14	413	0.16	0.31	749
Bengali	0.32	0.55	0.48	689	0.34	0.58	0.49	647	0.42	0.6	547
Chinese	0.06	0.34	0.24	358	0.09	0.31	0.22	395	0.16	0.48	643
Gujarati	0.17	0.33	0.26	164	0.15	0.39	0.24	122	0.19	0.38	143
Japanese	0.37	0.65	0.57	675	0.45	0.69	0.58	621	0.35	0.62	535
Modern Hebrew	0.16	0.3	0.23	97	0.28	0.65	0.51	96	0.25	0.55	155
Panjabi	0.21	0.43	0.35	372	0.3	0.48	0.4	321	0.23	0.48	323
Polish	-	-	-	44	0.24	0.48	0.43	78	0.18	0.38	137
Russian	0.24	0.54	0.49	495	0.22	0.52	0.42	503	0.25	0.6	646
Turkish	0.22	0.32	0.28	413	0.22	0.37	0.33	487	0.22	0.38	397
Urdu	0.42	0.58	0.52	2507	0.35	0.55	0.46	2281	0.34	0.53	2230
Persian	0.37	0.48	0.5	66	0.5	0.55	0.61	72	0.41	0.28	54
Classical Civilisation	0.51	0.78	0.65	1665	0.5	0.77	0.64	1788	0.56	0.82	2511
Classical Greek	0.2	0.72	0.46	87	0.14	0.77	0.37	117	0.25	0.76	426
Latin	0.32	0.76	0.6	2476	0.34	0.76	0.54	2395	0.36	0.79	4598
Music	0.55	0.71	0.65	38343	0.54	0.7	0.64	36635	0.56	0.71	37206
Physical Education/Sports Studies	0.56	0.71	0.66	105240	0.55	0.71	0.66	95437	0.53	0.69	86040
Dance	0.48	0.65	0.56	13910	0.46	0.64	0.55	13321	0.44	0.6	11494
Catering Studies	0.5	0.7	0.6	10230	0.49	0.71	0.61	12823	-	-	-
Office Technology	0.61	0.77	0.73	26247	0.6	0.77	0.72	21280	0.62	0.79	13246
General Studies	0.68	0.83	0.8	5305	0.67	0.83	0.8	6094	0.65	0.82	5321
D&T Electronic Products	0.51	0.74	0.65	10576	0.52	0.74	0.66	9838	0.52	0.75	8973
D&T Food Technology	0.55	0.76	0.67	59939	0.54	0.76	0.66	56933	0.56	0.77	49452
D&T Graphic Products	0.48	0.7	0.6	49823	0.48	0.7	0.6	46057	0.54	0.75	39360
D&T Resistant Materials	0.47	0.7	0.6	61639	0.47	0.69	0.6	56578	0.53	0.74	50754
D&T Textiles Technology	0.54	0.75	0.66	35284	0.55	0.76	0.67	32194	0.57	0.77	30643
D&T Systems &	0.47	0.68	0.6	5176	0.46	0.68	0.59	4920	0.52	0.73	3942

	Correlations with KS2, KS3 and mean concurrent GCSE (excluding given subject)										
	2009				2010				2011		
Subject	**KS2**	**Mean GCSE**	**KS3**	**N**	**KS2**	**Mean GCSE**	**KS3**	**N**	**KS2**	**Mean GCSE**	**N**
Control											
D&T Engineering	0.41	0.71	0.54	572	0.36	0.58	0.46	577	0.49	0.73	404
D&T Product Design	0.5	0.72	0.62	22102	0.49	0.71	0.62	25504	0.52	0.74	27974

Figure A1 below shows the relationship between the correlation between KS2 and GCSE grade found in 2010 and that found in 2011. As can be seen, these correlations are relatively stable over time. The GCSE subjects with the highest correlation with KS2 in 2010 also tend to be the ones with the highest correlation in 2011.

Figure A1: Correlations between GCSE subjects grades and KS2 in 2010 and 2011