



CAMBRIDGE ASSESSMENT

Analysis of use of Key Stage 2 data in GCSE predictions


Tom Benton and Tom Sutch

ARD Research Division

Final Report – 18th February 2014

Ofqual/14/5471

**Colour printing of this
report is strongly advised**

 UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

Ofqual
.....

This report has been commissioned by the Office of Qualifications and Examinations Regulation.

Table of Contents

1. Introduction	4 -
1.1 Comparable outcomes and GCSE predictions	4 -
1.2 Existing research	5 -
1.3 Aims of current project	7 -
1.4 Data	8 -
1.4.1 Data provided by awarding organisations (AOs)	8 -
1.4.2 Data from the National Pupil Database	8 -
2. Review of current method of generating predictions	10 -
2.1 Description of current method	10 -
2.2 Possible alternative measures of KS2 attainment	12 -
2.3 Correlations between different measures of KS2 attainment and GCSE grades	16 -
2.4 Examining differences in KS2-GCSE correlations across subjects	20 -
2.5 Predictive power of different KS2 measures across years	24 -
2.6 Differences with predictions from screening (concurrent attainment)	28 -
2.7 Practical differences between predictions based on different measures	31 -
2.7.1 Further exploration of the effect of the KS2 grade inflation adjustment	36 -
2.8 Summary	43 -
3. Review of tolerances for reporting outcomes that do not meet predictions	44 -
3.1 Method and results	45 -
3.2 Comparison with tolerances calculated using simple random sampling (SRS) methods ..	48 -
3.3 Quantifying tolerances as percentage rather than percentage point changes	49 -
3.4 Expected difference with screening predictions	50 -
3.4 Summary	51 -
4. Review of differences between screening outcomes and predictions	52 -
4.1 Comparison of KS2 and screening predictions	52 -
4.2 Are screening predictions influenced by the combination of GCSE specifications - candidates have taken at GCSE?	54 -
4.3 Possible solutions to the issue of under-prediction of AO differences	56 -
4.3.1 Adjusting the KS2 method using ideas from equating	56 -
4.3.2 Controlling for centre-level attainment in predictions	57 -
4.3.3 Using historical differences to adjust predictions	57 -
4.4 Final thoughts on the under-prediction problem	64 -
4.5 Summary	65 -
5. Appropriate tolerances for predictions based on concurrent GCSE performance	66 -
5.1 Comparison with tolerances calculated using simple random sampling (SRS) methods ..	67 -
5.2 Summary	68 -
6. Differences in the relationship between KS2 and GCSE achievement between years - and AOs	69 -
6.1 Differences between years	69 -
6.2 Differences between AOs	73 -
6.3 Summary	76 -
7 Further investigation of centre effects	77 -
7.1 Using centre type in predictions	77 -
7.1.1 Is the separate treatment of candidates from selective and independent schools - justified?	79 -
7.1.2 Does accounting for centre type in the model give any benefit?	80 -
7.2 Controlling for mean centre-level KS2	82 -
7.3 Summary	82 -
8. Further work and final thoughts	83 -
8.1 Summary of results	83 -

8.2 Other issues not explored	84 -
8.3 Final note.....	85 -
References.....	87 -
Appendix 1: GCSE predictions using mean Key Stage 2 Level as the measure of prior - attainment.....	89 -
Appendix 2: Detailed description of methodology used to estimate tolerances for each AO - and each subject	95 -
Further validation of the method	96 -
Appendix 3: A modified method for producing GCSE predictions based upon Key Stage 2 -	99 -
Appendix 4: Examination of the relationship between KS2 match rate and agreement of - results with screening outcomes.....	100 -

1. Introduction

To help ensure that GCSE and A level results are comparable with the standards of previous years, awarding organisations (AOs) use data on pupil attainment to predict the percentage of candidates expected to achieve the key grades (such as GCSE grades A*, A and C) in each subject overall. This is a key tool for guiding awarders when they set grade boundaries and for maintaining standards over time.

To predict the expected outcomes for any given year's GCSE cohort, AOs look at the relationship between GCSE performance in a relevant reference year and that cohort's attainment at Key Stage 2 (KS2) (where available). This allows them to produce a model of the relationship they can use to produce expected outcomes for the given year's GCSE cohort. A detailed description of the process used for the majority of predictions in 2013 is given in Appendix 1. A more general description of the process is provided within Section 2.

The aim of the research in this report is to provide a thorough technical evaluation of the relationship between GCSE results and prior attainment at KS2, including a consideration of whether predictions can be made more valid, and a review of the general approach in terms of using KS2 data to support the maintenance of standards. This report will also examine the continuing validity of using average KS2 attainment to produce predictions given that the last national KS2 Science tests took place in 2009, and hence no data from these tests will be available for the 16 year old GCSE cohort of 2015.

1.1 Comparable outcomes and GCSE predictions

The use of GCSE predictions based on KS2 attainment to help define GCSE grade boundaries is part of Ofqual's wider strategy known as *comparable outcomes*. This means that, under usual circumstances¹, the aim is that "... roughly the same proportion of students will achieve each grade as in the previous year." (Ofqual, 2012, page 2)².

The aim to achieve comparable outcomes is explicitly set against the aim for each grade to represent *comparable performance* over time. On the one hand this is argued for from the basis of avoiding a dip in grades whenever a new specification is introduced as teachers become used to the new material. However, it is also explicitly intended to combat "grade inflation". That is, the focus on comparable outcomes is intended to reduce the extent to which there are increases in the percentage of students achieving the highest grades year-on-year.

Given the overarching aim to ensure that the overall grade distribution will be roughly equivalent between years, the next task is to decide upon how grades should be distributed across different subjects and (within subjects) across different AOs. A simple approach might be for each AO to simply award the same number of GCSEs at each grade in each subject as they did in the previous year. However, whilst such an approach may be acceptable as a rough rule of thumb, it fails to take account of the fact that centres may switch which AO they enter their candidates with in any subject and so both the number and the nature of the candidates entering with each AO will change over time. Equally it may be that certain subjects become more popular as a whole with different types of candidates over time. In either case, it is desirable for the way in which grades are distributed between subjects and between AOs to be able to account for such changes. Furthermore, given that the aim is to explicitly focus on comparable outcomes rather than comparable performance, it is clear that statistical predictions will be at the heart of the process, with examiners responsible for ensuring the statistically recommended grade boundaries are appropriate.

¹ See page 3 of Ofqual (2012) for exceptions.

² This is not the only possible interpretation of the phrase "comparable outcomes". Elsewhere, it can be interpreted instead as being an outworking of the Similar Cohort Adage (Newton, 2011) where it is assumed that if the characteristics of two cohorts (such as groups of candidates studying with two different AOs) are very similar then their GCSE pass rate (at any grade) should also be similar.

It is well understood that the best predictor of a candidate's future attainment is their prior attainment (Benton, Hutchison, Schagen, and Scott, 2003), Figures 38 and 39). Furthermore, since 2011, the only widely available measure of prior attainment is provided by the results of national testing at KS2. For this reason it is natural that any statistical method to produce predictions of likely outcomes for different AOs and different subjects should focus upon attainment at KS2.

1.2 Existing research

Several existing research studies examine the relationship between prior attainment at KS2 and subsequent GCSE achievement. In the context of GCSE awarding, Eason (2010) examined the effectiveness of using KS2 to predict GCSE achievement for various AQA specifications. Predictions based upon KS2 were compared to predictions derived using concurrent GCSE attainment; a measure of a candidate's total achievement across all subjects rather than just the one of interest. The results of this analysis were promising in that for 89 per cent of the 168 grade boundaries analysed (each of grades A*, A, C and F across each of 42 subjects) the KS2-based predictions were within +/-1 per cent of those based on concurrent GCSE performance. Similar analysis by Benton and Sutch (2012) likewise found that KS2-based predictions tended to be close to those derived from mean GCSE.

Outside of the context of GCSE awarding, KS2 data is used widely to predict the likely performance, and hence set targets for individual candidates as part of the 'RAISEonline system' (Association of School and College Leaders [ASCL] 2011). To support this use of KS2 data, work by Treadaway (2013) compared the predictive power of several measures of prior attainment including Cognitive Ability Tests (CATs) and MidYis³ assessments taken in Year 7 to the predictive power of KS2. His results showed that KS2 achievement was generally more strongly correlated with achievement at either KS3 or GCSE than either CATs or the MidYis assessments. However, these findings relied upon KS2 being quantified in terms of sub-levels and the analysis showed that the correlations were very slightly lower if average KS2 levels were used instead⁴. For this reason, the 'RAISEonline system' uses sub-levels to produce its predictions.

Further analysis of the effect of different ways of quantifying KS2 achievement, within the context of GCSE awarding was undertaken by Eason 2010 (examining created prior attainment deciles based upon total KS2 raw scores) and Eason 2012 (which also examined the use of normalised KS2 scores⁵). This research also suggests that predictions based upon normalised KS2 scores (converted into deciles) may provide more accurate predictions than the current approach based on KS2 levels. More detailed analysis of the impact of using different measures of KS2 to create predictions will be provided in Section 2.

On a more negative note Smith (2013) examined the strength of the relationship between KS2 achievement and GCSE grades in individual subjects. This analysis noted that the strength of the association was small in absolute terms⁶, particularly for Modern Languages and for practical subjects, although stronger relationships were found in the core subjects of English, Mathematics and Science. The report also raises concerns about the way in which KS2-based predictions are adjusted for grade inflation in KS2 itself. This issue will be discussed more thoroughly in Section 2.7.1.

³ Middle Years Information System Tests provided by the Centre for Evaluation and Monitoring (CEM) at Durham University.

⁴ Depending on which outcome was analysed correlations with average KS2 *levels* were occasionally slightly lower than correlations with CATs but never lower than correlations with MidYis assessments.

⁵ These will be described in more detail in the next section.

⁶ Figures from this report are presented as pseudo-R square coefficients rather than correlations. In addition to this several different coefficients are presented so there is no one single figure that can be quoted. However the report states (page 7) that "at best, we can estimate that around 38% of the variation in GCSE grade can be predicted by KS2 category and some measures were considerably less than this".

The extent to which the relationship between KS2 and GCSE is stable between different centre types was examined by Eason 2010. This research suggested that KS2 under-predicts likely attainment within independent and selective centres. For this reason, it recommended that these centre types are excluded from predictions; an approach that has currently been adopted as standard practice. Issues relating to accounting for different centre types within predictions will be explored further in the final report due to be completed in January 2014.

Some research has examined the effectiveness of predictions based upon *common centres* as an alternative to KS2. That is, assuming that, as a group, centres that enter candidates for the same subject in successive years should achieve similar results. Although previous research has suggested this method has some validity (Eason 2009; Benton and Sutch 2012) the analysis by Benton and Sutch indicated that such predictions were further from the “gold standard” predictions based on concurrent attainment than predictions from KS2⁷. Other research (Eason 2003, 2006) suggests reasons for caution in using predictions based upon common centres. For example, individual centres may split their GCSE entries between different specifications according to ability thus affecting the validity of common centre predictions. Furthermore, it is clear that large changes in the size of GCSE entries within any subject are commonplace within common centres. This implies that we cannot guarantee that the candidates entering a GCSE subject within a centre one year are comparable to the candidates entering that subject within the same centre the following year.

Several previous studies attempt to examine the expected reliability of KS2-based predictions of GCSE outcomes for individual AOs and subject. Some early work by Pinot de Moira (2008), based upon statistical modelling, suggested that the level of reliability is more dependent upon the proximity of the prediction to 50 per cent and the size of the entry for a given award than the correlation between prior attainment and outcomes. However, the estimates in this report failed to account for the clustering of candidates within centres and the impact of centres on the results of individual pupils. Further work by Benton and Lin (2011) used a more complex non-parametric technique to estimate the reliability of predictions at AS and A level based upon prior attainment at GCSE. This work has been used to produce *tolerances* for predictions at GCSE (based on KS2) as well as AS and A levels; that is, guidance as to how closely awards by each AO should match with predictions before an explanatory report is required (Ofqual, 2013). However, similar calculations to those of Benton and Lin (Benton and Sutch, 2012) examining the reliability of GCSE predictions based upon prior attainment at KS2, have suggested that the derived tolerances may be too tight at grade C. The issue of the reliability of KS2-based predictions will be examined further in Section 3.

Other analysis (Smith, 2013) compares the size of currently recommended tolerances with the width of 95 per cent statistical confidence intervals for simple random samples of different sizes and suggests that current tolerances are too small. There are various reasons why the estimated reliability of KS2-based predictions does not need to necessarily match with expectations based upon simple random sampling (clustering of candidates within centres, the fact that estimates are derived for a fixed level of prior attainment). However, further analysis of the reliability of KS2-based predictions (Section 3) will also examine the relationship between properly calculated confidence intervals and those generated using statistical formulae for simple random samples.

Several of the studies above compare predictions based upon KS2 to predictions based upon concurrent attainment (Benton and Sutch 2012; Eason 2010, 2012). Although intuitively appealing due to the high correlation between attainment in one GCSE subject and achievement in others, a potential problem with this approach is that each method may be fundamentally

⁷ Furthermore, further recent analysis of this same data suggests that this is not only caused by the fact that using a common centres technique implies aiming for a different standard overall, but also because there is greater variability in the predictions based on common centres than in predictions based upon KS2. Both techniques are seeking to estimate the same quantity; a comparable outcome for GCSEs. However, this is done less reliably using data from common centres than by using KS2.

applying a different standard. Predictions of national achievement rates in any subject based on concurrent GCSE attainment assume that achievement should remain relatively constant in the population of candidates taking GCSEs. In contrast, predictions based on KS2 assume that achievement should remain relatively constant amongst the population of candidates that were entered for KS2. Alternatively, predictions based upon English and Mathematics GCSEs only (Spalding, 2012, Unpublished) assume that national outcomes in each subject should remain the same for the population of candidates taking both English and Mathematics GCSE. However, it is worth noting that the populations for whom achievement is assumed to be fixed are not exactly the same across the different methods. For example, not all pupils that take KS2 will go on to take GCSE – they may take alternative qualifications such as IGCSEs or the International Baccalaureate instead. Similarly, some pupils within the GCSE population will not have valid KS2 results available. Thus, assuming that GCSE outcomes would be fixed for one of these populations is not the same as assuming it would be fixed for another. Thus, the different approaches are inherently aiming to maintain slightly different standards, and, for this reason, it is difficult to discern the extent to which the results of these studies, and the slight differences between predictions that are found, reveal information about the accuracy of different methods or the similarity of the underlying assumptions. To avoid this issue, our own analysis of the comparison between predictions using concurrent attainment and those using KS2 (Section 2.8 and Section 4) will focus on predictions of the extent to which outcomes for different AOs are predicted to be above or below the national average in each subject. That is, it will focus on relative, rather than absolute, predictions for each subject and each AO. This approach will be discussed further in section 2.6.

1.3 Aims of current project

This report examines the following issues:

- Whether the accuracy of GCSE predictions could be improved by quantifying - achievement at KS2 differently? (Section 2). -
- What are the most appropriate tolerances for GCSE predictions? That is, how closely should we expect the outcomes awarded by AOs within any subject to match predictions? (Section 3).
- How do predictions from KS2 compare to predictions from concurrent GCSE attainment and what are the reasons for any differences? (Section 4).
- What would be the most appropriate tolerances for GCSE predictions if a method based upon concurrent attainment was used to produce these? (Section 5).
- How does the relationship between KS2 and GCSE achievement vary over time and between different AOs? (Section 6).
- What difference would it make if centre type were also accounted for within models as well as KS2 attainment? (Section 7).

The recommendations from the various analyses will be summarised in Section 8 which will also provide details of further issues arising from the use of statistical predictions in GCSE awarding.

1.4 Data

The data used in this project was provided from two different sources:

1.4.1 Data provided by awarding organisations (AOs)

The AOs⁸ provided data sets detailing the performance of 16 year old candidates in each of their GCSEs in each of June 2009-2013 inclusive. This included matching information regarding the average KS2 level for each candidate. Achievements in any GCSE specification were grouped into subjects according to the categorisation used by AOs to produce prediction matrices⁹. Any achievements in qualifications that were not included in this list were removed from the data¹⁰.

This data is used for the majority of analysis in Sections 3 and 4.

1.4.2 Data from the National Pupil Database

The AO data described in Section 1.4.1 contains KS2 levels for candidates entering GCSEs, as currently used to generate predictions and guide awarding, but none of the more detailed data necessary to explore alternative measures of KS2 attainment as described in Section 2. As a substitute, we used the Key Stage 4 (KS4) tables from the National Pupil Database (NPD), which is maintained by the Department for Education (DfE) and covers pupils in England only. These tables contain information at an individual candidate level on the results of GCSEs (among other qualifications), along with pre-matched demographic and prior attainment data, including levels and raw marks from KS2. Five years of NPD data were used from 2009–2013 inclusive.

This data differs very slightly from that provided by AOs in that it only includes candidates who were studying in England¹¹. However, both sets of data were restricted to candidates certificating in the June sessions, both sets of data made use of *all* of a candidate's achievement in any subject (rather than restricting to their best grade in any subject), and both sets of data are restricted to GCSEs achieved in Year 11 (that is, 16 year old candidates)¹². Furthermore, GCSEs in this data set were only included in analysis if they were amongst the list of specifications grouped into subjects by AOs by agreement through JCQ.

In order to replicate the data used for GCSE predictions, the dataset for each year was restricted to results for GCSEs taken in the summer series by 16 year olds. Only GCSEs awarded by AQA, Edexcel, OCR or WJEC were kept for analysis. If a candidate had more than one GCSE result recorded for the same specification with an AO, the highest was taken; however, no de-duplication was performed across different awarding bodies (that is, candidates entering GCSE Mathematics with both AQA and Edexcel would have both present), as would occur with the predictions.

Along with information on GCSEs (subject, AO, specification number) and KS2 attainment (levels and raw marks), demographic information was extracted from the NPD. Individual candidates' gender and level of deprivation, as measured by the Income Deprivation Affecting Children Index (IDACI), and characteristics of the centre through which the candidate was

⁸ Specifically AQA, Edexcel, OCR and WJEC.

⁹ This is the same list that is agreed by Ofqual and used for all live analysis for GCSE awarding. The only difference for the purposes of our analysis here is that, in order to simplify the results of analysis, each GCSE specification was assigned to exactly one subject. This implied that: we only considered maths as a whole subject rather than also producing separate predictions for linear and modular maths, *combined* English Language and Literature qualifications were counted as English, and individual modern foreign languages were always analysed separately rather than as part of any overall MFL grouping.

¹⁰ A brief inspection of the specifications removed in this way revealed they tended to be non-GCSEs. GCSE short courses were also removed from analysis as part of this process.

¹¹ Although for the purposes of this research, the requirement for candidates to have completed KS2 almost restricts the data to England anyway, a small number of cross-border candidates may occur in the AO data that are not included within the NPD.

¹² Thus any candidates resitting a GCSE will be included in analysis but that grades achieved in any early entries will not be counted.

entered for examination: centre type, the local authority area and hence region in which it was located was also included.

The KS2 data in Section 1.4.1 was used to determine candidates' normalised KS2 scores, and which quantiles the candidates fell into (using a variety of different measures). Fine grades at KS2, and hence sub-levels, were calculated using standard methods (see DfE, 2011, Annex D, and ASCL, 2011). GCSE results with a missing grade (for example 'X', indicating absence) were excluded. Mean GCSE score (taking U=0, G=1, F=2, ..., A=7, A*=8) was calculated for pupils with at least three GCSE results, and this was used to create deciles. All analysis for Section 2 was carried out for GCSE entries (within subjects for each candidate) for which the full set of predictors was available (KS2 and GCSE mean decile).

This data was used for any analyses requiring more detailed information on achievement at KS2 including all analyses reported in Section 2, and some of the analysis in Section 4.

2. Review of current method of generating predictions

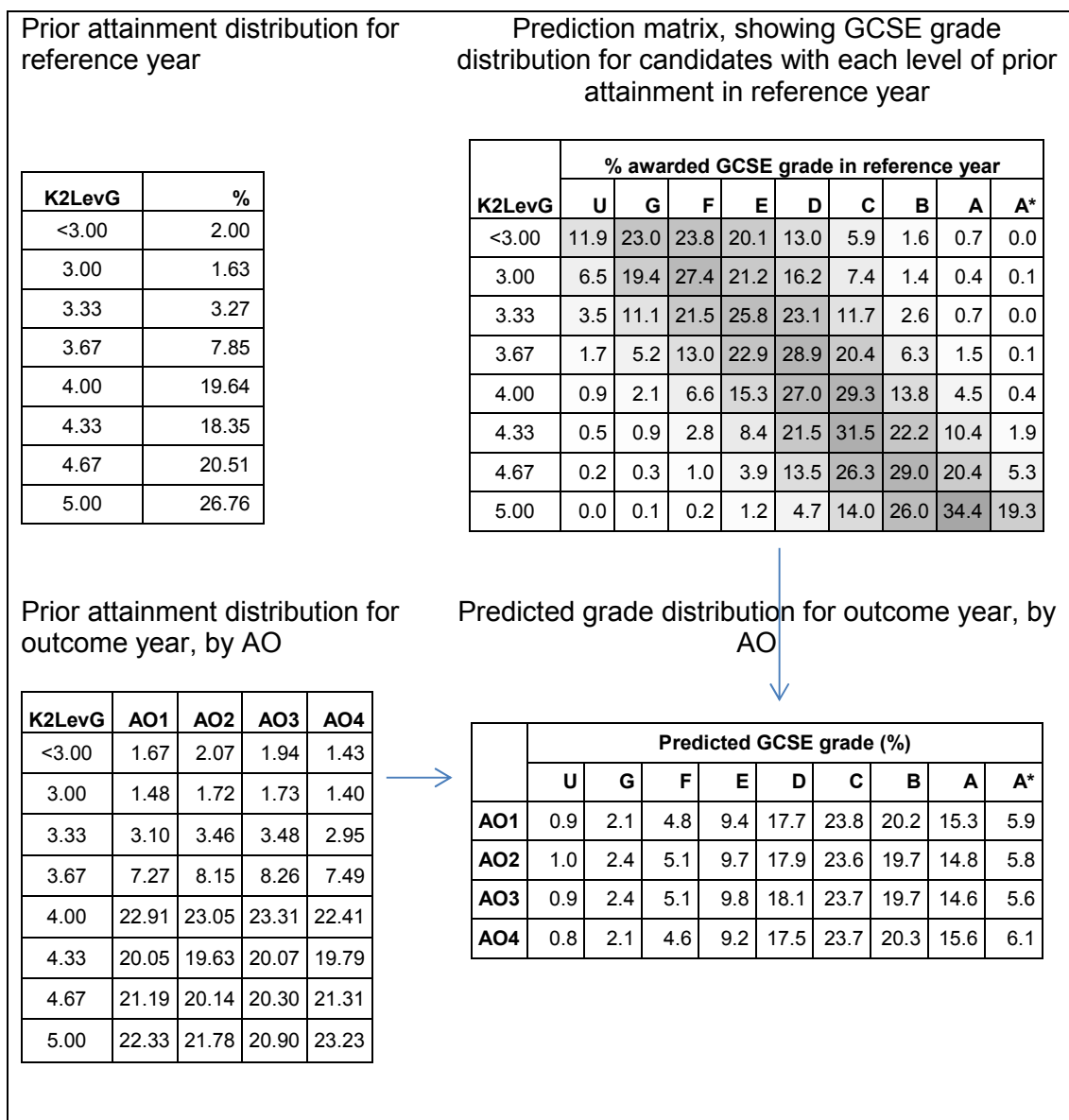
This section gives an overview of the current method of generating predictions at GCSE using prior attainment at KS2, and explores a number of alternative measures, evaluating each against several criteria.

2.1 Description of current method

At its simplest, a prediction matrix for a given GCSE subject requires data on prior attainment (at KS2) and outcomes (GCSE grades) for candidates taking GCSEs in that subject with all awarding bodies in a given year, termed the *reference year*. This matrix is then used by each AO to predict outcomes in the current year for the cohort of candidates taking its exams with known prior attainment. This approach allows for the ability of the cohort taking a particular subject with a particular AO to vary from year-to-year when predictions are calculated, but with the underlying assumption that the relationship between prior attainment and outcomes remains constant. Because of differences evident in the *value added* between KS2 and GCSE, candidates from independent and selective schools are excluded from the prediction matrix (see Eason 2010). It is important to note that the predictions are intended to guide the awarding at the cohort level, and not predict the grades that individual candidates will receive.

In the current method, each candidate's KS2 attainment is calculated as a simple mean of their levels in English, Mathematics and Science. Only candidates with recorded levels in each of three subjects are included in analysis. Each subject level can take the value B (below the level of the tests), N (no test level awarded), 2, 3, 4 or 5, or other indicators of missing data, for example indicating absence or an inability to access the tests. When they are averaged, B and N are set to zero but other missing levels are excluded from the calculation. Candidates are categorised according to their average level, and those with an average of less than 3 are grouped together. This gives eight possible categories: less than 3, 3.00, 3.33, 3.67, 4.00, 4.33, 4.67 and 5.00. The outcomes for each of these categories are computed for the reference year (for example, 31.5 per cent of candidates with prior attainment of 4.33 gained a C in a particular subject) and applied in the current year to each AO's cohort of candidates in the subject. Figure 2.1 shows an example.

Figure 2.1 Example showing application of simple prediction matrix, without adjustment for KS2 inflation



The assumption of a constant relationship between prior attainment and GCSE grade will break down if there is a change in the standard associated with a given level of prior attainment. One potential cause of this change might be grade “inflation” or “deflation” at KS2. Furthermore, because the stated aim of comparable outcomes is that “roughly the same proportion of students will achieve each grade as in the previous year” (Ofqual 2012, page 2) it is intended that *national* gains in achievement at KS2 do not necessarily translate into increases in national results at GCSE. In order to address this, a post-hoc adjustment is applied by calculating the predictions that would result if the whole national cohort of KS2 candidates, in each of the reference and outcome¹³ years, had entered the GCSE in the given subject (for any AO). Assuming that the ability of the whole cohort at KS2 is constant from one year to the next, the difference between these two predictions at each grade is used to adjust the raw predictions obtained above. The implications of this adjustment are explored further in Section 2.7.1.

¹³ That is, the current year at the time of live calculation. Within the report, we will occasionally refer to these as the “outcome years” due to the fact that they do not now necessarily relate to the current year (that is, they are not always 2013 in our calculations).

The exact application of the method is somewhat more complex than outlined above. Due to changes in specifications, Ofqual's data exchange procedures for each year state which year (or years) should be used as the basis of predictions (the reference year). Where more than one year is used, an average of the predictions for each grade is generated, weighted by the entry in each year¹⁴. The approach, including grade inflation adjustment and allowing for the use of multiple years to guide predictions, is explained fully in Appendix 1.

2.2 Possible alternative measures of KS2 attainment

There are a number of potential issues with the current measure of KS2 attainment: it is coarse (based on whole levels within each subject), susceptible to any grade inflation at KS2, and although there are eight possible categories, candidates are predominantly bunched in the top four of them. As such it may be possible to improve on the accuracy of predictions by using a slightly different measure. In addition, KS2 tests in Science for all candidates were discontinued in 2010 (having been replaced by a sampling system to monitor national standards) so it is necessary to consider the effect of excluding KS2 Science for all alternative measures. The last cohort to have sat national KS2 tests in Science will be in Year 11 in 2014.

The measures considered are as follows:

- K2LevG: mean KS2 level in English, Mathematics and Science (weighted equally), as used currently. There are eight possible categories, ranging from 2.67 (which contains all values of less than 3.00) to 5.00. The distribution of this measure, shown in Table 2.1, is negatively skewed: 70–75 per cent of candidates are bunched in the top four categories, which may impede discrimination at the top.
- K2LevGEM: mean KS2 level in English and Mathematics only (weighted equally). The distribution of this measure is shown in Table 2.2. There are six possible categories, ranging from 2.5 (which contains all values of less than 3.0) to 5.0. As with K2LevG, there is bunching in the top three categories.
- K2SubG: sub-level in English, Mathematics and Science. Fine point scores are calculated for each of the KS2 subjects, using standard methodology (DfE, 2011), which effectively interpolates within levels based on raw marks. A mean is then taken (with equal weighting between the three subjects) to give one fine point score per candidate. Sub-level groups are calculated as in ASCL documentation (ASCL, 2011), giving ten categories. The distribution, shown in Table 2.3¹⁵, is still negatively skewed, but there is slightly more discrimination at the top end as very few candidates receive the highest sub-level of 5a.
- K2SubGEM: sub-level in English and Mathematics only. This measure is calculated in the same way as K2SubG, but excludes Science marks. The distribution is shown in Table 2.4 and exhibits similar properties to K2SubG.
- K2RwTo. This measure is a simple total of raw marks in English, Mathematics and Science tests where the raw scores have been imputed, if necessary, to the middle of the mark range of each level for the small number of candidates with available test levels but without any raw marks recorded. Furthermore all candidates with a test level of B (that is, below the level of the test) were assigned a raw score of zero. This measure implicitly gives a lower weighting to Science than K2LevG and K2SubG as the maximum mark in the Science KS2 paper is 80, rather than 100 as in English and Mathematics. This measure is more susceptible to variations in the demand of individual KS2 papers

¹⁴ For example, in 2013, for GCSE specifications first certificated in summer 2011, data from GCSEs in 2011 and 2012 was used. But for new Science GCSE specifications that certificated for the first time in 2013, predictions were generated using 2011 data only.

¹⁵ It is important to note that due to differences in the way the two are calculated, the proportion of candidates achieving sub-levels 5c, 5b and 5a is substantially different for the proportion of candidates with an average level of 5.

between years and subjects than are levels, which are adjusted through the level setting process according to the demand of the paper, or sub-levels. The distribution, shown in Figure 2.2, is negatively skewed but with a modal mark of zero. It is clear that the distribution varies between years.

- K2RwToEM. This measure is a simple total of raw marks in English and Mathematics KS2 tests (including imputations as described above). As with K2RwTo, this measure is more susceptible to variations in demand of individual KS2 papers, and the distribution, shown in Figure 2.2, is negatively skewed but with a modal mark of zero.
- K2NrTo. To compute this measure, the percentile rank of each candidate in each KS2 subject was calculated and converted to the equivalent point on a normal distribution with a mean of 50 marks and standard deviation of 16.67 marks¹⁶. The three normalised marks were then summed (thereby giving equal weight to English, Mathematics and Science). As Figure 2.2 shows, this has removed the negative skew of the distribution but there is still a 'spike' corresponding to candidates with zero raw marks.
- K2NrToEM. This measure was calculated in the same way as K2NrTo, but excluding normalised Science marks from the total. The distribution, shown in Figure 2.2, shows similar properties to that of K2NrTo.
- Finally, candidates were assigned to octiles, deciles and quindeciles (15 categories) based on each of K2RwTo, K2RwToEM, K2NrTo and K2NrToEM. Quantiles based on raw marks are denoted as K2RwG8, K2RwG10, K2RwG15, K2RwG8EM, K2RwG10EM and K2RwG15EM and a similar convention is followed for quantiles based on normalised marks. Octiles were selected because there are eight categories used in the current measure (K2LevG) – albeit distributed unevenly – so this measure will help determine whether simply having a more even spread of candidates improves prediction. Deciles were selected as they are used to predict A level based on mean GCSE score (Benton and Lin, 2011). Quindeciles are to see if extra granularity improves prediction.

This gives a total of 20 measures of KS2 attainment. Many of these measures are susceptible to KS2 grade inflation in the same way as the current method based on K2LevG (this applies to K2LevGEM, K2SubG, K2SubGEM, K2RwTo, K2RwToEM), and hence a similar post-hoc adjustment has been used to adjust for them. For quantiles and normalised marks, no adjustment is necessary.

The total raw marks and normalised marks (K2RwTo, K2RwToEM, K2NrTo, K2NrToEM) were used to fit multinomial logistic regression models, which model the probability that a candidate with a given KS2 mark m would achieve GCSE grade i (where $0=U$, $1=G$, ..., $7=A$, $8=A^*$) in a particular subject as:

$$\log\left(\frac{p_i}{p_8}\right) = \beta_{0i} + \beta_{1i}m \quad (0 \leq i \leq 7)$$

Note that the probabilities are used with respect to a reference category, in this case 8 (an A* grade). The requirement that $\sum_{i=0}^8 p_i = 1$ allows the model to be uniquely specified.

For the other 16 measures, along with deciles based on mean concurrent GCSE score, a simple matrix-based approach was used (akin to the current method described in Section 2.1).

¹⁶ This matches the approach of Eason (2012). It also ensures that total normalised scores cover roughly the same range as total raw scores (see Figure 2.2).

Table 2.1: Distribution of K2LevG

	KS2 Year				
	2003/4	2004/5	2005/6	2006/7	2007/8
<3.00	8.08	7.63	7.55	7.04	6.57
3.00	3.72	3.67	3.68	3.38	3.15
3.33	6.57	6.15	5.96	5.53	5.35
3.67	11.30	10.62	10.61	10.44	10.15
4.00	20.99	20.28	19.77	20.44	22.95
4.33	16.05	17.28	15.57	15.89	16.96
4.67	15.67	16.96	15.78	16.01	16.33
5.00	17.61	17.41	21.07	21.28	18.55

Table 2.2 Distribution of K2LevGEM

	KS2 Year				
	2003/4	2004/5	2005/6	2006/7	2007/8
<3.0	8.38	7.83	7.78	7.25	6.74
3.0	7.77	7.53	7.24	6.79	6.38
3.5	13.86	13.39	13.13	12.79	12.18
4.0	30.53	31.84	28.78	29.69	33.08
4.5	20.58	21.07	20.67	20.98	21.78
5.0	18.88	18.34	22.40	22.51	19.84

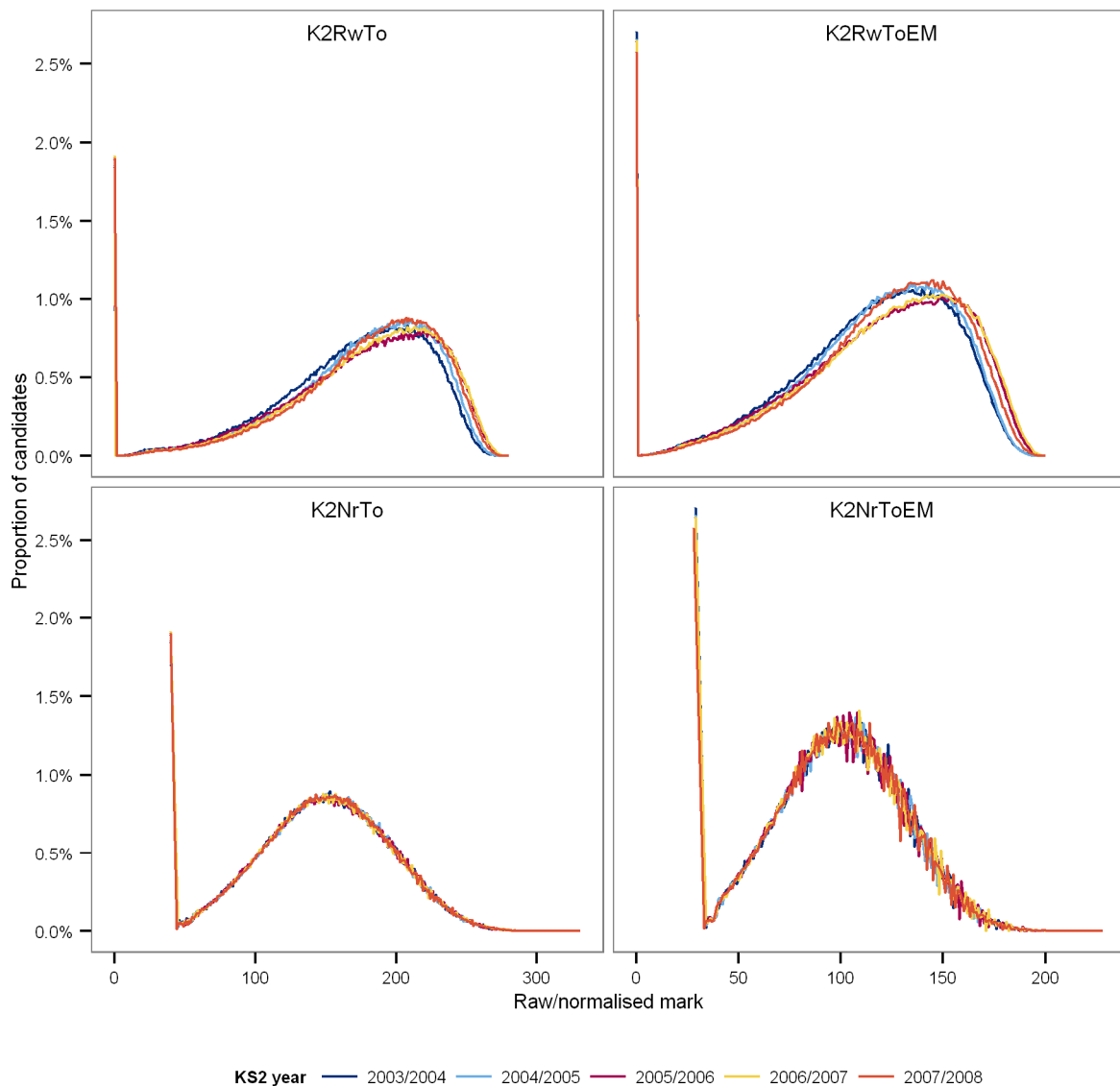
Table 2.3: Distribution of K2SubG

	KS2 Year				
	2003/4	2004/5	2005/6	2006/7	2007/8
2	4.22	4.10	3.97	3.79	3.52
3c	3.02	2.77	2.84	2.49	2.23
3b	4.87	4.56	4.62	4.17	3.75
3a	7.96	7.43	7.40	7.02	6.56
4c	12.54	11.90	11.53	11.40	11.25
4b	17.68	17.52	16.50	16.96	17.96
4a	20.21	20.71	19.45	20.44	21.85
5c	18.47	19.34	18.92	19.42	19.99
5b	10.43	10.95	13.14	12.66	11.76
5a	0.60	0.73	1.62	1.66	1.15

Table 2.4: Distribution of K2SubGEM -

	KS2 Year				
	2003/4	2004/5	2005/6	2006/7	2007/8
2	5.94	5.51	5.45	5.06	4.64
3c	3.65	3.43	3.43	3.05	2.69
3b	5.55	5.33	5.26	4.96	4.50
3a	8.83	8.50	8.07	7.81	7.33
4c	13.53	13.58	12.50	12.53	12.56
4b	17.85	18.44	16.83	17.33	18.60
4a	18.56	19.43	18.33	19.14	20.87
5c	16.06	16.75	17.30	17.11	18.03
5b	9.31	8.52	11.34	11.27	9.84
5a	0.72	0.52	1.48	1.74	0.95

Figure 2.2: Distributions of raw and normalised marks



2.3 Correlations between different measures of KS2 attainment and GCSE grades

The correlation between any of the KS2 measures and the grade achieved in each GCSE subject provides an indication of the accuracy of predictions within a given year, and allows us to discount the effect of inter-year variation in candidature and any inflation at KS2. Although the outcomes and, in most cases, the predictors are discrete rather than continuous, Pearson correlations are presented here in order to facilitate comparison with other studies and to provide more familiarity to readers¹⁷.

The correlations have been calculated for each GCSE subject, excluding candidates from selective and independent schools, and only including candidates for whom all potential KS2 predictors are available (for example, raw marks at KS2 as well as levels) and who took at least three GCSEs.

Table 2.5 shows the median subject-level correlation¹⁸ of each KS2 predictor with GCSE grade in each year, for subjects with entries of 400 or above¹⁹. Correlation with deciles based on concurrent mean GCSE is also shown. There is very little difference between the correlations using KS2 predictors: all of them are around 0.5 and are markedly lower than correlation with concurrent GCSE (which is approximately 0.8²⁰). The current method (K2LevG) and K2LevGEM had among the lowest correlations with GCSE grade, while the highest correlations were found for raw and normalised marks (K2RwTo, K2NrTo), and in some years by K2NrG15 and K2RwG15. The correlations have been fairly stable between years, but with a slight tendency to increase over time (this is also evident for concurrent deciles based on mean GCSE).

In most cases, the more fine-grained predictors have a very slightly stronger correlation with GCSE grade than the coarser predictors. For example, the median correlation for total raw marks (K2RwTo) is typically²¹ greater than for 15 categories (K2RwG15) which is in turn greater than for 10 and 8 categories (K2RwG10 and K2RwG8), although these differences are typically less than 0.01. A similar pattern is evident for the predictors based on normalised marks.

Correlations for predictors based on English and Mathematics at KS2 were slightly lower (by approximately 0.01) than for predictors based on English, Mathematics and Science.

Correlations for normalised marks (K2NrTo) were higher than for raw marks (K2RwTo), and there were slightly higher correlations for the quantiles based on normalised marks (K2NrG15, K2NrG10, K2NrG8) than the corresponding quantiles based on raw marks.

¹⁷ Pearson correlations are, technically, most appropriate for use with continuous variables.

¹⁸ That is, the median of the correlations calculated for each of the GCSE subjects.

¹⁹ The number of subjects included ranged from 55 in 2011 to 59 in 2012 and 2013.

²⁰ Similar correlations between GCSE grades in individual subjects and concurrent attainment were identified by Benton and Sutch (2012). This earlier research yielded high correlations despite the fact that, for each GCSE subject being studied, the mean GCSE grade was calculated based upon each candidate's achievement in all of their GCSEs in *other* subjects (i.e. not included the one for which the correlation is being calculated). This indicates that the inclusion of the subject for which the correlation is being calculated in the measure of mean GCSE does not have a large influence upon the results presented here.

²¹ With the exception of 2013, where the median correlation with K2RwTo was very slightly lower than that with K2RwG15.

Table 2.5: Median correlation with GCSE grade for subjects with entry of at least 400 candidates

Predictor variable	Median subject-level correlation with GCSE grade				
	2009	2010	2011	2012	2013
K2LevG	0.505	0.484	0.515	0.503	0.506
K2LevGEM	0.489	0.480	0.506	0.498	0.491
K2SubG	0.508	0.506	0.528	0.517	0.521
K2SubGEM	0.505	0.496	0.522	0.516	0.510
K2RwTo	0.517	0.511	0.535	0.529	0.528
K2RwToEM	0.500	0.504	0.530	0.526	0.522
K2RwG8	0.512	0.500	0.529	0.522	0.521
K2RwG8EM	0.494	0.497	0.521	0.520	0.511
K2RwG10	0.514	0.504	0.530	0.527	0.525
K2RwG10EM	0.496	0.499	0.523	0.520	0.513
K2RwG15	0.517	0.506	0.534	0.529	0.529
K2RwG15EM	0.499	0.501	0.526	0.524	0.518
K2NrTo	0.526	0.523	0.550	0.538	0.539
K2NrToEM	0.514	0.514	0.538	0.537	0.529
K2NrG8	0.516	0.506	0.534	0.523	0.524
K2NrG8EM	0.501	0.505	0.524	0.523	0.513
K2NrG10	0.518	0.510	0.538	0.525	0.529
K2NrG10EM	0.507	0.507	0.526	0.526	0.516
K2NrG15	0.520	0.512	0.541	0.529	0.530
K2NrG15EM	0.514	0.510	0.528	0.529	0.519
Meangc10	0.794	0.798	0.803	0.803	0.808

Figure 2.3 shows the distribution of correlations for individual subjects for each predictor (from which the subject-level medians in Table 2.5 are drawn). All correlations are positive but there are some outliers with very low correlations; similarly some subjects have a very high correlation (around 0.75). These will be examined in detail in Section 2.4.

Figure 2.3: Subject-level correlation with GCSE grade by KS2 predictor, for subjects with entry of at least 400 candidates

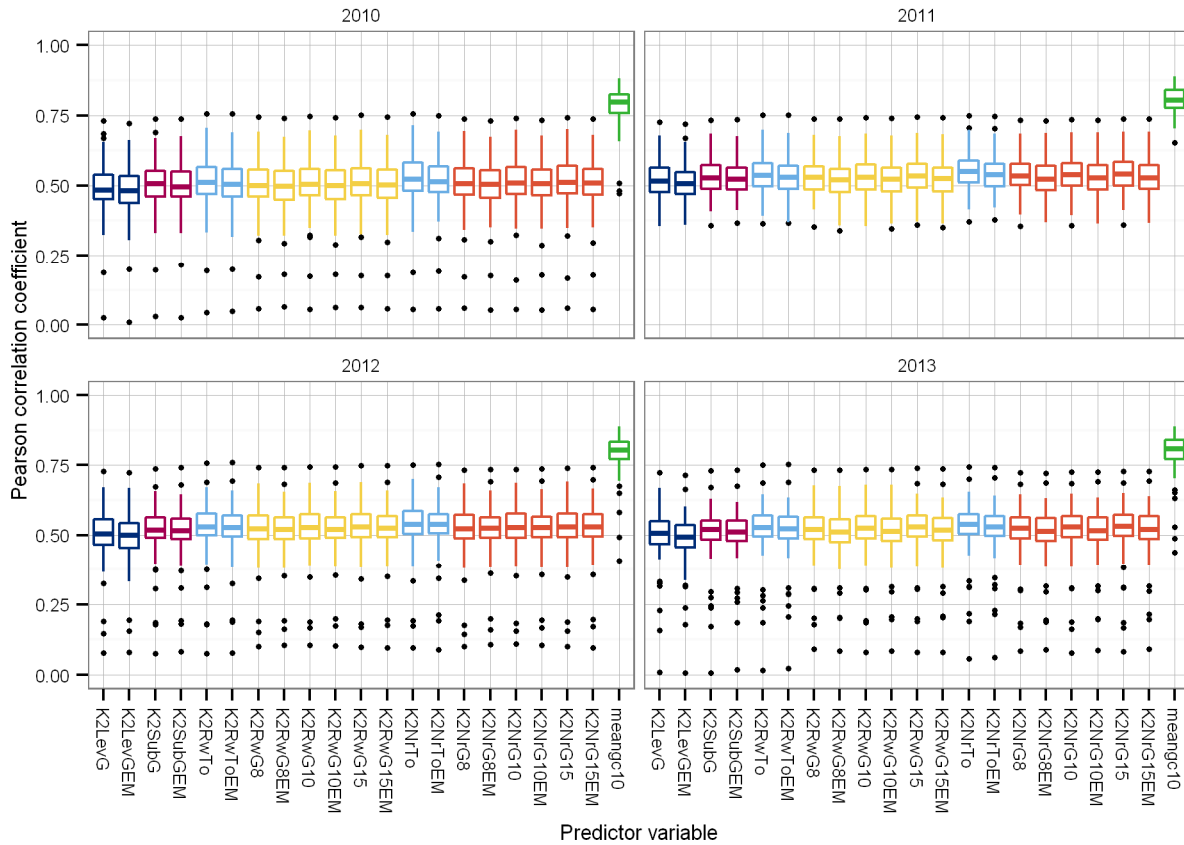
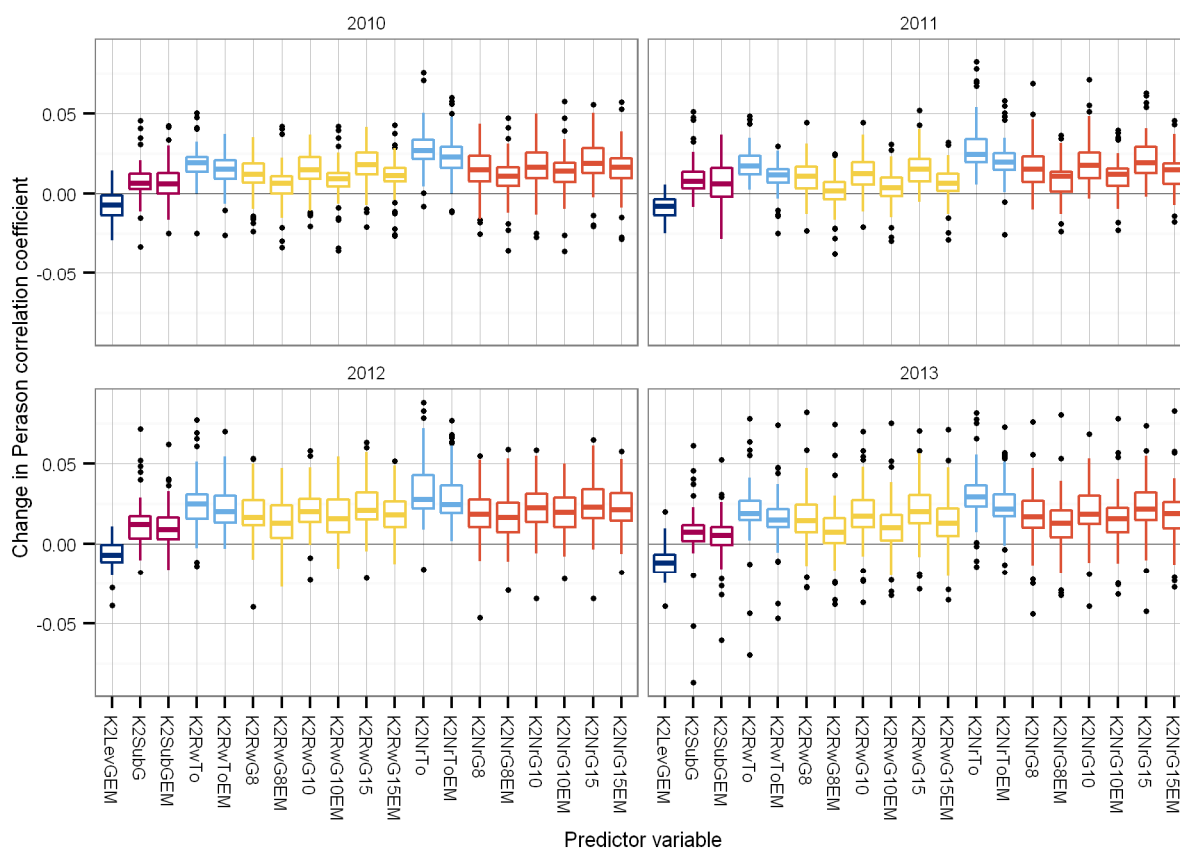


Table 2.6 and Figure 2.4 present the improvement in correlation for each of the KS2 predictors compared to the current method (K2LevG). It is clear that potential gains in correlation are small (the highest being K2NrTo, at just below 0.03). Moving to levels based on English and Mathematics alone, (K2LevGEM) would reduce correlations by approximately 0.01.

Table 2.6: Median improvement in correlation compared to current method (K2LevG) for subjects with entry of at least 400 candidates

Predictor variable	Median subject-level improvement in Pearson correlation coefficient				
	2009	2010	2011	2012	2013
K2LevGEM	-0.011	-0.007	-0.008	-0.007	-0.012
K2SubG	0.006	0.006	0.008	0.012	0.007
K2SubGEM	0.000	0.006	0.006	0.009	0.005
K2RwTo	0.016	0.019	0.017	0.025	0.019
K2RwToEM	0.004	0.015	0.012	0.020	0.015
K2RwG8	0.007	0.012	0.011	0.016	0.015
K2RwG8EM	-0.004	0.007	0.002	0.013	0.007
K2RwG10	0.010	0.015	0.013	0.020	0.017
K2RwG10EM	-0.003	0.009	0.004	0.015	0.010
K2RwG15	0.012	0.018	0.015	0.021	0.020
K2RwG15EM	0.000	0.011	0.007	0.018	0.013
K2NrTo	0.025	0.027	0.024	0.028	0.029
K2NrToEM	0.014	0.023	0.020	0.025	0.022
K2NrG8	0.015	0.015	0.015	0.018	0.017
K2NrG8EM	0.003	0.011	0.011	0.017	0.013
K2NrG10	0.016	0.017	0.018	0.022	0.018
K2NrG10EM	0.005	0.014	0.012	0.020	0.016
K2NrG15	0.019	0.019	0.019	0.023	0.022
K2NrG15EM	0.008	0.016	0.015	0.021	0.019

Figure 2.4: Subject-level improvement in correlation with GCSE grade compared to current method, for subjects with entry of at least 400 candidates



2.4 Examining differences in KS2-GCSE correlations across subjects

Table 2.7 shows the correlation between selected KS2 predictors and GCSE grade for each subject in 2013. The highest correlations with KS2 predictors can be found in Mathematics and English (these are also closest to the correlation with concurrent attainment). In previous years, Single Science also had a high correlation (0.700 in 2009) but this has fallen away, probably due to changes in patterns of entry (Single Science is now generally entered in Year 10, and 15 year olds are excluded from standard prediction matrix calculations). This is unsurprising given that these are the subjects assessed at KS2. However, it is notable that Biology, Chemistry and Physics have rather lower correlations with KS2 attainment, probably due to the nature of the entry (predominantly high ability candidates).

The lowest correlations are with minority Modern Languages (Arabic, Mandarin, Turkish, Urdu and Bengali), which may be taken by native speakers who achieve higher grades in their first language than in other subjects, and Applied Art & Design. These subjects also have the lowest correlation with concurrent attainment.

In contrast to prior attainment, the subjects which have the highest correlation with concurrent attainment are Geography and History. So concurrent attainment appears to be measuring something different to KS2, and, furthermore, as seen from the earlier correlations, this different metric is more closely related to achievement in individual GCSEs subjects.

As previously shown in Figure 2.4, in most subjects and years there would be an improvement by switching to another measure (except K2LevGEM). The subjects where alternative measures have the greatest improvement on correlations are Latin, the three Separate Sciences, French, Astronomy, and Economics, where gains of around 0.03–0.06 could be obtained by moving to sub-levels instead, and gains of around 0.05–0.08 could be gained by using normalised scores.

These subjects are taken disproportionately by high ability candidates, and this effect is likely to be because the current measure lacks discrimination at the top end, but other measures such as sub-levels are able to give finer detail.

At the other extreme, the correlation with sub-level is lower for General Studies, Bengali, Urdu and Arabic and a number of applied subjects: Home Economics (Food & Nutrition), Citizenship Studies, Health & Social Care, Engineering, and Applied Art & Design). There is a particular deterioration in correlation when moving to sub-levels or raw marks for Bengali, Applied Art & Design and Urdu.

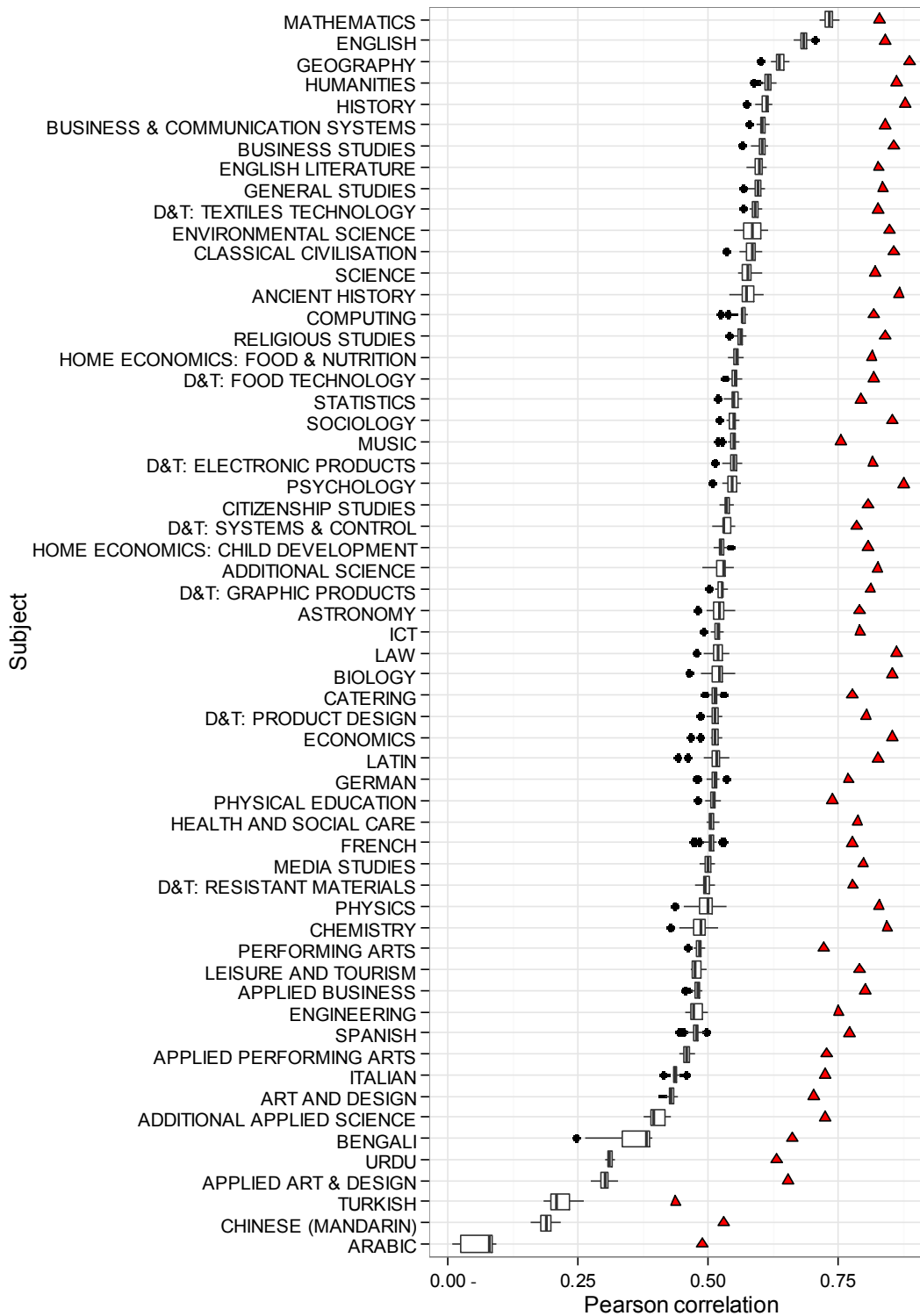
Table 2.7: Correlations between selected KS2 measures and GCSE grade, for 2013 (highest average KS2 correlations at the top)

<i>Subject</i>	<i>n</i>	<i>Correlation with GCSE grade</i>					
		<i>K2LevG</i>	<i>K2LevGEM</i>	<i>K2SubG</i>	<i>K2RwTo</i>	<i>K2NrTo</i>	<i>meangc10</i>
MATHEMATICS	366710	0.722	0.714	0.731	0.751	0.743	0.828
ENGLISH	443596	0.669	0.665	0.670	0.686	0.700	0.840
GEOGRAPHY	154118	0.621	0.601	0.630	0.645	0.655	0.888
HUMANITIES	6347	0.595	0.589	0.599	0.614	0.627	0.862
HISTORY	185349	0.589	0.573	0.601	0.615	0.623	0.878
BUSINESS & COMMUNICATION SYSTEMS	8543	0.593	0.580	0.601	0.615	0.618	0.841
BUSINESS STUDIES	50753	0.583	0.565	0.597	0.614	0.615	0.857
ENGLISH LITERATURE	350650	0.581	0.573	0.589	0.603	0.610	0.828
GENERAL STUDIES	3954	0.577	0.567	0.576	0.592	0.610	0.836
D&T: TEXTILES TECHNOLOGY	23003	0.579	0.567	0.585	0.598	0.603	0.827
ENVIRONMENTAL SCIENCE	782	0.589	0.550	0.597	0.607	0.615	0.849
CLASSICAL CIVILISATION	906	0.558	0.534	0.574	0.587	0.593	0.858
SCIENCE	92387	0.576	0.556	0.587	0.599	0.603	0.821
ANCIENT HISTORY	638	0.560	0.541	0.561	0.589	0.605	0.867
COMPUTING	2937	0.538	0.524	0.550	0.571	0.575	0.817
RELIGIOUS STUDIES	180755	0.550	0.540	0.556	0.568	0.573	0.841
HOME ECONOMICS: FOOD & NUTRITION	7168	0.548	0.538	0.546	0.559	0.567	0.815
D&T: FOOD TECHNOLOGY	36226	0.542	0.532	0.544	0.556	0.565	0.819
STATISTICS	22758	0.529	0.518	0.544	0.562	0.566	0.794
SOCIOLOGY	16787	0.534	0.521	0.545	0.557	0.559	0.854
MUSIC	30197	0.528	0.520	0.542	0.553	0.559	0.756
D&T: ELECTRONIC PRODUCTS	6797	0.529	0.514	0.540	0.551	0.564	0.817
PSYCHOLOGY	9789	0.528	0.506	0.536	0.551	0.562	0.877
CITIZENSHIP STUDIES	8831	0.532	0.522	0.529	0.541	0.547	0.808
D&T: SYSTEMS & CONTROL	2783	0.531	0.508	0.533	0.540	0.553	0.785
HOME ECONOMICS: CHILD DEVELOPMENT	14368	0.520	0.511	0.524	0.538	0.542	0.807
ADDITIONAL SCIENCE	215414	0.513	0.489	0.531	0.543	0.550	0.826
D&T: GRAPHIC PRODUCTS	30112	0.512	0.501	0.516	0.527	0.538	0.812
ASTRONOMY	908	0.498	0.480	0.521	0.524	0.552	0.791
ICT	38233	0.506	0.491	0.517	0.528	0.531	0.792
LAW	1931	0.496	0.478	0.496	0.512	0.538	0.862
BIOLOGY	116154	0.485	0.463	0.522	0.540	0.551	0.855

<i>Subject</i>	<i>n</i>	<i>Correlation with GCSE grade</i>					
		<i>K2LevG</i>	<i>K2LevGEM</i>	<i>K2SubG</i>	<i>K2RwTo</i>	<i>K2NrTo</i>	<i>meangc10</i>
CATERING	20294	0.506	0.494	0.509	0.521	0.531	0.776
D&T: PRODUCT DESIGN	26349	0.497	0.485	0.502	0.514	0.527	0.803
ECONOMICS	2738	0.484	0.465	0.514	0.525	0.528	0.854
LATIN	1203	0.461	0.442	0.522	0.539	0.539	0.827
GERMAN	43570	0.479	0.478	0.495	0.513	0.535	0.770
PHYSICAL EDUCATION	74929	0.494	0.479	0.506	0.518	0.523	0.739
HEALTH AND SOCIAL CARE	6715	0.502	0.495	0.500	0.513	0.518	0.787
FRENCH	117016	0.474	0.472	0.484	0.501	0.527	0.778
MEDIA STUDIES	41085	0.485	0.482	0.490	0.502	0.509	0.798
D&T: RESISTANT MATERIALS	39817	0.488	0.473	0.488	0.501	0.513	0.778
PHYSICS	115953	0.453	0.435	0.499	0.517	0.535	0.828
CHEMISTRY	115923	0.443	0.427	0.484	0.502	0.519	0.844
PERFORMING ARTS	64690	0.471	0.462	0.479	0.489	0.493	0.722
LEISURE AND TOURISM	2107	0.483	0.469	0.485	0.490	0.490	0.791
APPLIED BUSINESS	3696	0.461	0.455	0.467	0.480	0.486	0.802
ENGINEERING	1793	0.477	0.455	0.471	0.479	0.498	0.751
SPANISH	58494	0.446	0.445	0.452	0.469	0.492	0.772
APPLIED PERFORMING ARTS	1801	0.444	0.443	0.452	0.463	0.467	0.728
ITALIAN	2132	0.413	0.419	0.414	0.432	0.452	0.724
ART AND DESIGN	125414	0.418	0.411	0.418	0.428	0.442	0.703
ADDITIONAL APPLIED SCIENCE	10492	0.414	0.397	0.420	0.427	0.426	0.725
BENGALI	568	0.333	0.337	0.246	0.263	0.334	0.662
URDU	2014	0.315	0.314	0.295	0.302	0.314	0.632
APPLIED ART & DESIGN	813	0.326	0.318	0.274	0.282	0.311	0.653
TURKISH	459	0.228	0.238	0.237	0.237	0.217	0.437
CHINESE (MANDARIN)	492	0.159	0.179	0.171	0.186	0.191	0.529
ARABIC	595	0.009	0.007	0.008	0.016	0.058	0.488

This is also illustrated in Figure 2.5, where the black boxes and points show the distribution of correlation of all KS2 measures, and the red triangles on the right of the plot are the correlation with mean concurrent GCSE grade.

Figure 2.5: Correlation of KS2 and mean GCSE measures with GCSE grade, for 2013



The stability of the correlations between years was investigated but is not presented here in detail. In general, correlations were similar from one year to the next, although there were notable exceptions: correlations in Science decreased from around 0.70 in 2009 to 0.54 in 2013, with the sharpest drop between 2012 and 2013. In this subject, entries also reduced substantially as the current pattern is for most candidates to take the exam in Year 10 (these

candidates are therefore not in our dataset). By contrast, in Latin the correlation steadily increased from 0.31 in 2009 to 0.51 in 2011 while the entry volume remained stable. However, the majority of candidates in Latin attend selective or independent schools and are hence excluded from our data.

2.5 Predictive power of different KS2 measures across years

Having investigated the within-year correlations for each of the KS2 measures, we now examine how predictive models constructed using each measure in one year perform at predicting grade distributions in a different year. For a model to have validity in predicting outcomes and thus guiding awarding, it is important that the embodied relationships are generalisable across years, rather than being over-fitted to a particular year.

The criterion we have used to assess predictive power is deviance - a statistical measure of model fit based on the likelihood of a given set of results under the prediction model. Lower deviances indicate a better model fit. Deviance is calculated at the individual candidate and GCSE subject level as minus 2 times the logarithm of the probability (under the prediction model) of a candidate achieving their actual grade, and is then summed across candidates. For predictors that require adjustment for KS2 inflation, the adjustment at a particular GCSE grade was applied for all values of the KS2 predictor.

If a candidate achieves a grade that has a predicted probability of zero (as would arise when no candidate in the reference year with equivalent prior attainment gained that particular GCSE grade), this would theoretically result in an infinite deviance. To avoid this, all probabilities were truncated to be in the range 0.001 to 0.999 before deviance was calculated (in line with Benton and Lin, 2011).

As total deviance is larger for subjects with larger entries, we divide the deviance in each subject by the total number of candidates, which allows inter-subject comparison and prevents the model fit results being dominated by the performance in subjects with large entries such as Mathematics and English.

One issue is that, as the predictions from KS2 level have been used to guide the awarding of GCSEs, the actual grade distribution would be expected to closely match the predicted distribution from this method. Whilst this does not guarantee that the current method will generate the smallest deviance, it may slightly bias results towards favouring the current method. In order to investigate this, average deviances per candidate were calculated for each GCSE grade in each subject, then reweighted according to the predicted grade distributions by each of the measures. It was found that this hardly affected deviances at all, either actual magnitude or rank order among predictors, and therefore for simplicity the actual grade distribution was used.²²

²² Indeed, before 2011 predictions were not based on KS2 data at all. The lack of a step change in deviances between 2010 and 2011 suggests that any bias is negligible.

The median of the subject-level average deviances under each method for 2013 (using 2012 as a reference year) is shown in Table 2.8 along with the median absolute and relative difference from the current method²³. The distribution of the difference in deviance across subjects is shown in Figure 2.6 (for predictions using the previous year as a reference year) which shows few changes between years²⁴.

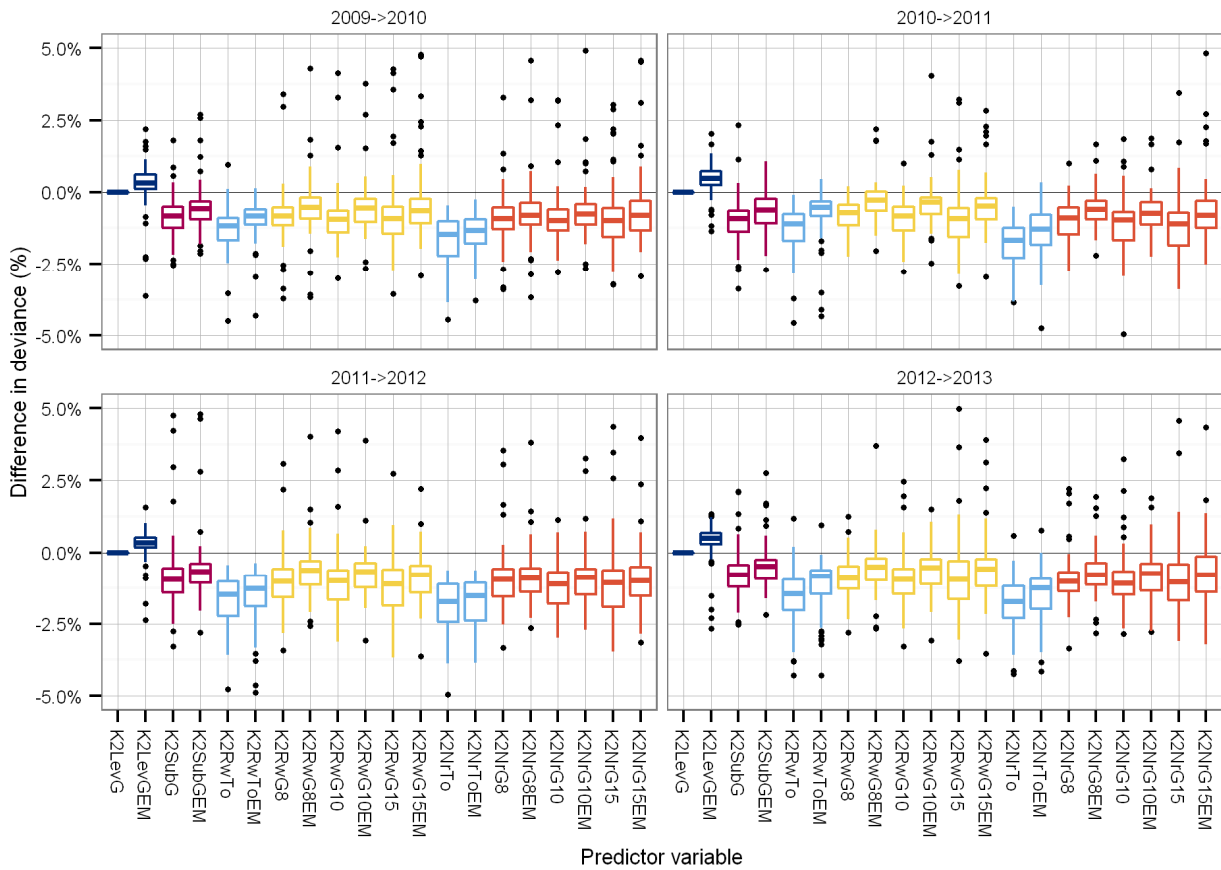
Table 2.8: Median deviance for each predictor (outcome year 2013, reference year 2012)

Predictor variable	Median value of average deviance per candidate	Median difference compared to current method (K2LevG)	Percentage difference compared to current method (K2LevG)
K2LevG	3.427		
K2LevGEM	3.438	+0.016	+0.50%
K2SubG	3.396	-0.028	-0.80%
K2SubGEM	3.406	-0.017	-0.49%
K2RwTo	3.384	-0.050	-1.45%
K2RwToEM	3.397	-0.030	-0.84%
K2RwG8	3.397	-0.030	-0.90%
K2RwG8EM	3.406	-0.018	-0.53%
K2RwG10	3.391	-0.034	-0.97%
K2RwG10EM	3.403	-0.019	-0.54%
K2RwG15	3.419	-0.032	-0.90%
K2RwG15EM	3.427	-0.019	-0.53%
K2NrTo	3.375	-0.058	-1.73%
K2NrToEM	3.380	-0.045	-1.25%
K2NrG8	3.390	-0.034	-1.02%
K2NrG8EM	3.395	-0.027	-0.81%
K2NrG10	3.387	-0.038	-1.09%
K2NrG10EM	3.387	-0.026	-0.74%
K2NrG15	3.427	-0.036	-1.04%
K2NrG15EM	3.423	-0.027	-0.75%
Meangc10	2.639	-0.793	-22.50%

²³ Note that, in general, the median of the differences is not equal to the difference of the medians, and (similarly) the median percentage difference is not equal to the percentage difference of the medians.

²⁴ The results in Figure 2.6 are based upon using a single reference year one year before the year in which predictions are made (e.g. using 2012 data to predict 2013). Further analysis was undertaken using a difference of two years between reference and outcomes years (e.g. using 2011 data to predict 2013). A very similar pattern of results was identified and so for brevity it is not included here.

Figure 2.6: Difference in deviance by using alternative measures



It is immediately clear that most other predictors give lower deviances than the current method, the exception being K2LevGEM in which Science KS2 results are excluded, although the reductions in deviances are fairly small. The greatest gain in predictive power comes from using normalised KS2 scores, with just a 1.7 per cent reduction in median deviance. This is contrasted to the reduction of more than 22 per cent that is achieved by using concurrent attainment, which suggests that any small gains in predictive power due to choosing one measure of KS2 rather than another are not worth pursuing. However, this issue will be returned to later once the practical differences between different sets of predictions have been explored.

Figure 2.7 presents the relationship between the deviance of each method and the number of candidates entering a subject. Each dot represents the deviance for a single subject (with all AOs combined) for a particular predictor variable, and each line is a smoothed mean. The overall trends evident from Figure 2.7 are that for the subjects with the highest entry, average deviance is lower and there is less variation in deviance between methods. However, there is a small but consistent reduction in deviance through using predictions based on total raw or normalised marks. Furthermore, this improvement is largest for subjects with low entry numbers.

Figure 2.7: Average deviance for selected KS2 measures, along with entry size, for 2012-2013

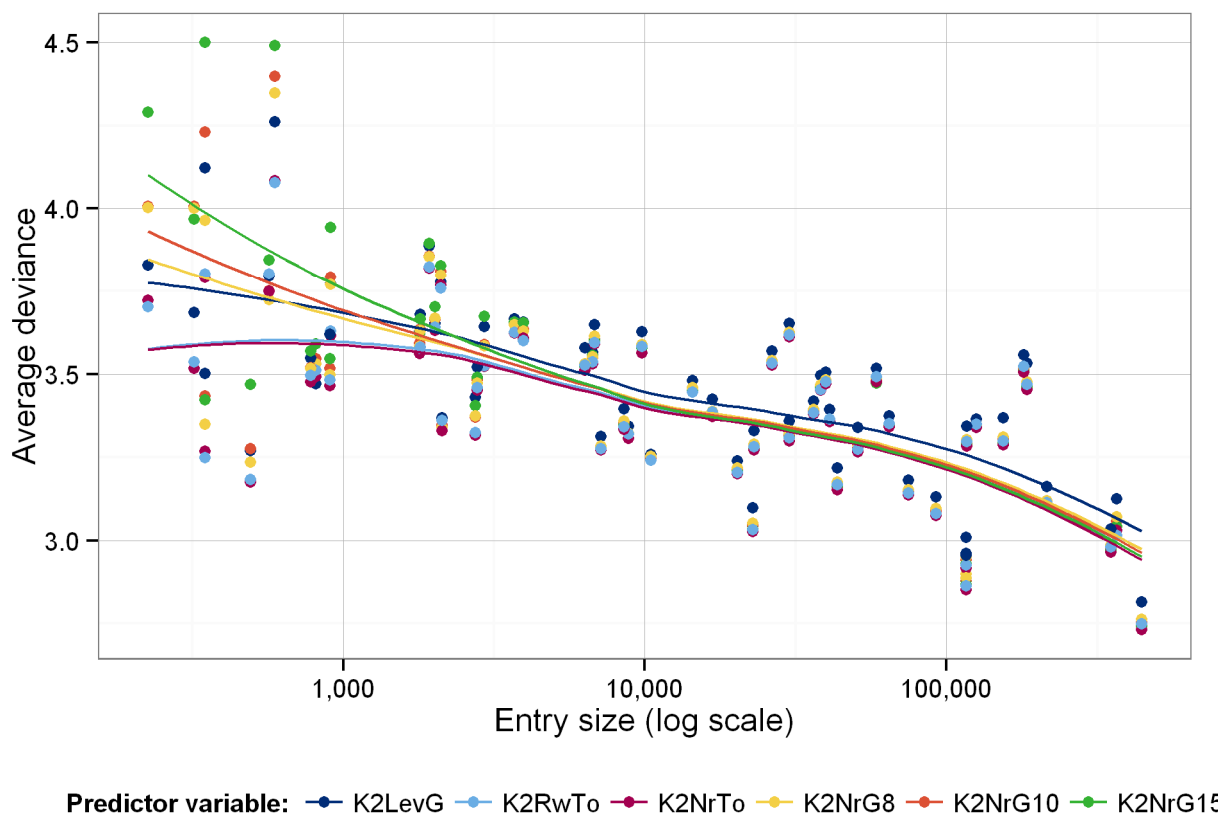


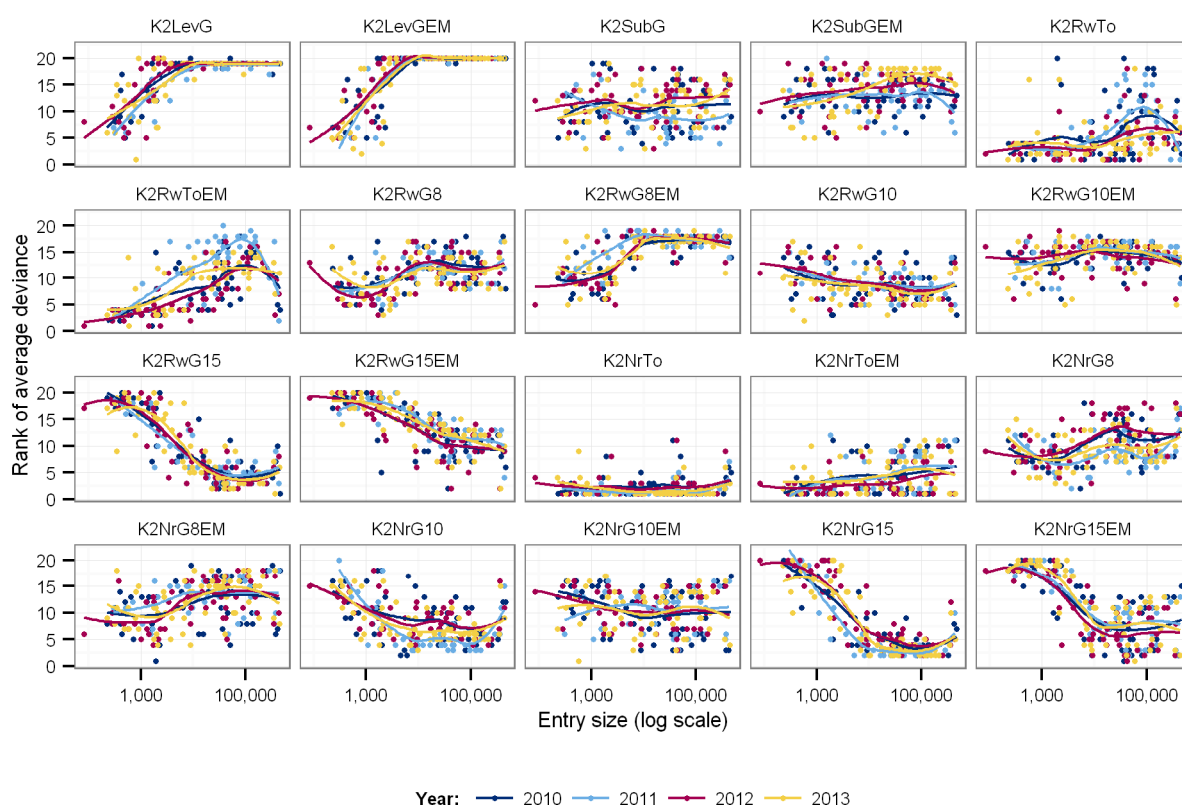
Figure 2.8 shows the relative ranking of the measures as entry size changes (where a rank of 1 indicates the best measure with the lowest deviance). There is a consistent pattern between years suggesting that the current method (K2LevG) performs relatively well for small entry subjects, but not for subjects with an entry of over 10,000. However, it should be remembered that the magnitude of the differences in deviance between the methods is smaller at this end too. In contrast, the reverse is true for methods based on quinciles (K2RwG15) which have the highest deviances for subjects with lower entry. This is an example of a phenomenon known as the *bias–variance trade-off*. When the number of candidates is small, the random variations between numbers of candidates in categories, and the relationship between prior attainment and GCSE grade, are a larger source of error than the bias implicit in a particular method (in this context, due to the degree of over-simplification of the underlying continuous relationship). The current method has high bias but low variance, while the quinciles have low bias but high variance.

The methods using logistic regression based on normalised total marks perform best consistently, no matter what the size of the entry is. For subjects with lower entry, it is advantageous that the estimates obtained via logistic regression are not too sensitive to small variations in normalised scores, whereas a small change in normalised score could have a big

effect if the candidate moves into a different quantile. Logistic regression using raw scores gives low deviances for low-entry subjects, but when the entry is higher, Figure 2.8 shows that it does not perform as well as the quantiles based on raw or normalised scores.

In view of the ‘spike’ evident in the distribution of total normalised score (K2NrTo) shown in Figure 2.2, corresponding to candidates with zero raw KS2 score, and also because of the relatively good predictive performance of K2NrTo, we also investigated accounting for these candidates separately by means of dummy variables in the logistic regression. However, we found that this had a negligible effect in practice (predictions were almost all within 0.1 percentage points) and so this possibility is not discussed further.

Figure 2.8: Rank order of subject-level deviances compared to entry size (predictions from consecutive years) 2010-2013



2.6 Differences with predictions from screening (concurrent attainment)

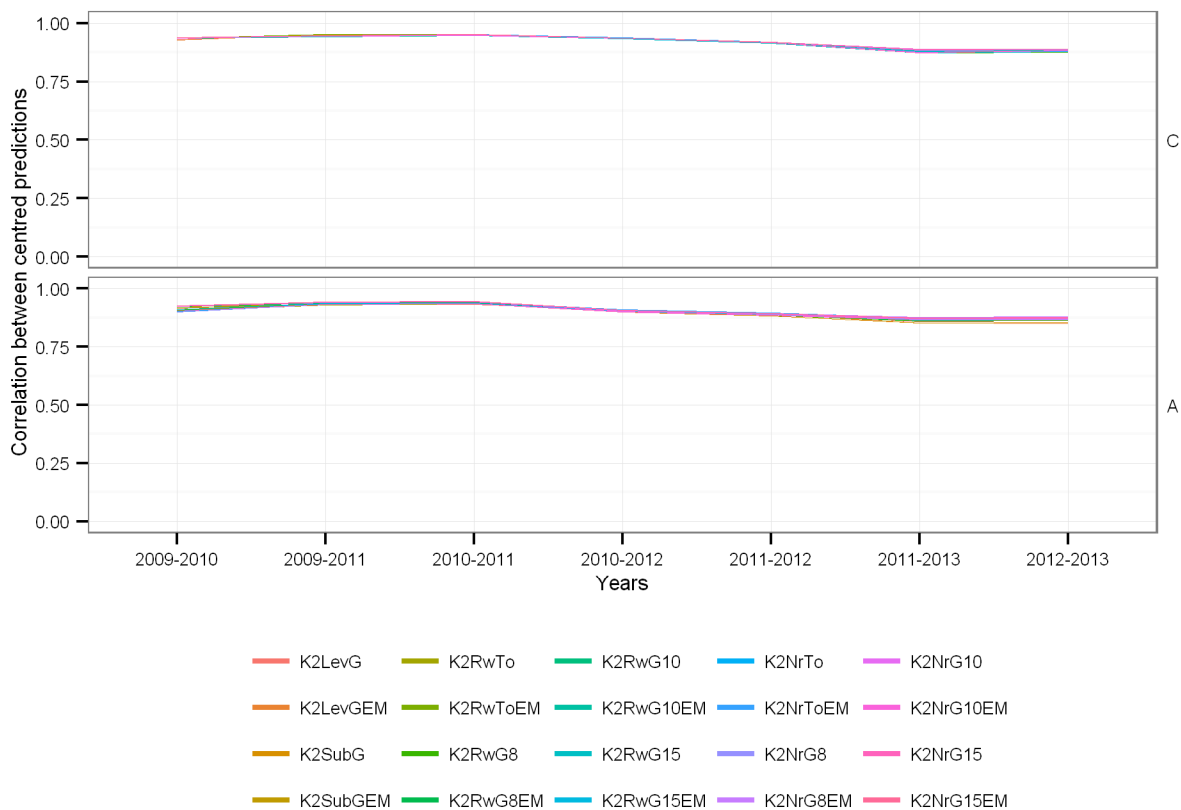
At face value, having consistency between the different methods used to ensure comparability between AOs is desirable. The annual inter-board screening exercise, in which awarding bodies carry out a statistical review of outcomes in each subject, in conjunction with candidates’ concurrent attainment, determines whether outcomes are comparable between AOs and flags subjects where one or more AOs are significantly out of line. In this section we make use of this ‘gold standard’ of differences in predictions made using concurrent attainment by comparing it with predictions made from prior attainment.

As discussed in Section 1.2, predictions using concurrent attainment assume consistency of outcomes amongst a different overall population than predictions based upon KS2. For this reason, rather than directly comparing the two sets of predictions, we examine *centred* predictions. That is, the difference between the percentage predicted to achieve a given grade or

above within a particular AO and the national percentage (across all AOs) predicted to achieve that or above. Clearly if a GCSE subject is only offered by a single AO then their prediction will equal the national prediction so that this approach is not possible. For this reason, such subjects have therefore been excluded from this analysis. Similarly, if a single AO has many more candidates in a given subject than any other AOs then it is virtually certain that both sets of predictions will lie close to the predicted national average. Such cases are included in analysis, but, since their centred predictions (both from KS2 and concurrent attainment) will be very close to zero, they will have very little effect on estimated correlations (see below) and upon the visual examination of differences. As such they do not prevent the identification of important differences between the two sets of predictions.

Centred predictions obtained from prior and concurrent attainment are very strongly correlated indicating that AOs with high ability candidates by one measure strongly tend to have high ability candidates by the other measure. Figure 2.9 shows that correlations between these measures are around 0.9, although reducing slightly over time. In addition, there is very little difference between the various KS2 measures. For example, in predictions obtained for 2013 using 2012 as a reference year, correlations ranged from 0.877 to 0.886.

Figure 2.9: Correlations between centred predictions from KS2 and concurrent attainment between 2009 and 2013



Scatterplots comparing the centred predictions for 2013 at grades C and A, using 2012 as a reference year, are presented in Figures 2.10 and 2.11. Each dot represents one subject from a particular AO. The correlations between the centred predictions from prior and concurrent attainment are very high for all KS2 measures. This indicates that the higher the difference in prior attainment between AOs, the higher the difference in concurrent attainment.

However, the scatter is off-diagonal: the magnitude of the centred prediction from KS2 tends to be less than the magnitude of the centred prediction from concurrent attainment. That is, KS2

tends to under-predict differences between AOs, or, putting it another way, the relationship with AOs is consistent, but under-represented by KS2. This will be explored further in Section 4.

Figure 2.10: Comparisons between centred predictions based on KS2 and concurrent attainment between 2012 and 2013 for grade C

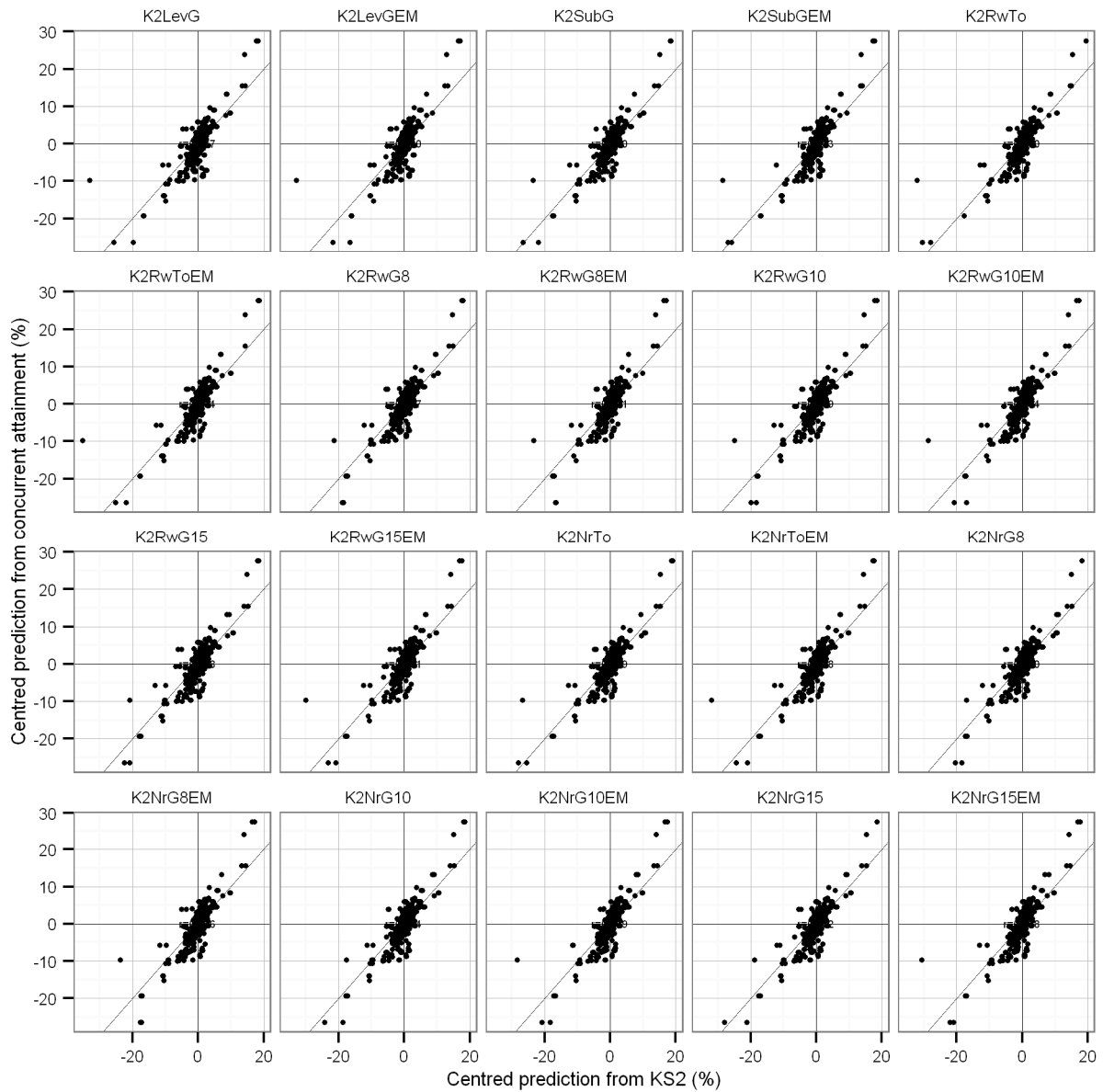
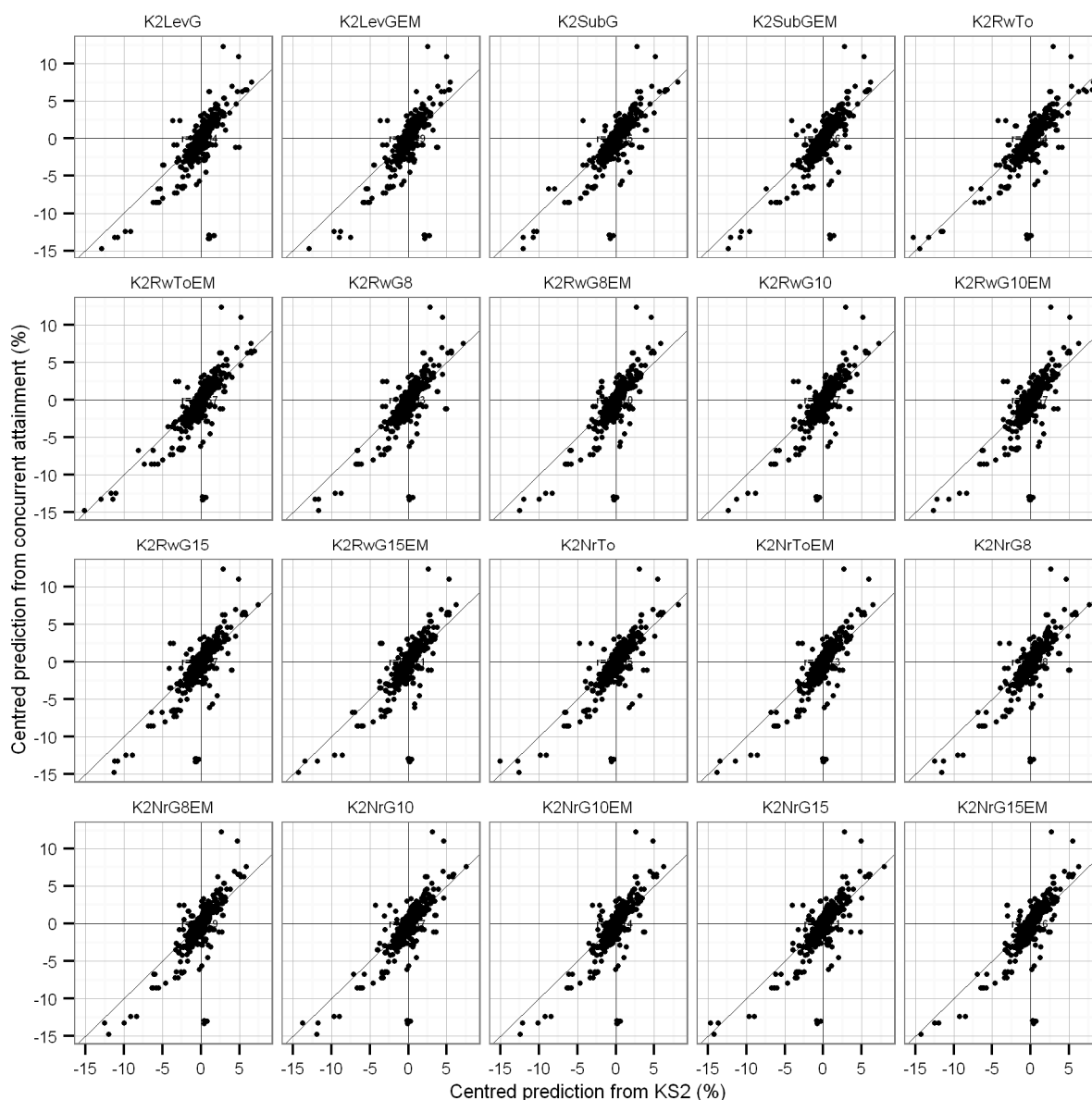


Figure 2.11: Comparisons between centred predictions based on KS2 and concurrent attainment between 2012 and 2013 for grade A



2.7 Practical differences between predictions based on different measures

The preceding sections have examined the relative performance of each measure of KS2 attainment, and established that some measures may give slightly more valid predictions. However, it is also of interest to determine how these methods would affect the predictions actually generated. As predictions are intended only as a guide for awarding, with awarding bodies subject to specified tolerances for reporting outcomes, a minor improvement in accuracy (of, say, 0.1 percentage points) will have little practical effect on the awarding process. It is also instructive to determine whether certain methods tend to result in predictions that are consistently more lenient or harsher than the current methods.

Figures 2.12–2.14 show the resulting differences for each set of reference and application years at grades C, F and A respectively, while Table 2.9 presents the median and interquartile range of difference for 2013 predictions only, using 2012 as a reference year.

On the whole, differences in predictions from the current method are very small. Even for grade C, where the largest differences arise as it is nearer the middle of the distribution, differences as

large as 1 percentage point are rare. For most predictors, the zero line (representing no difference) lies between the lower and upper quartiles.

From the boxplots illustrated in Figures 2.12-2.14, it is clear that there have been differences in the pattern over time. Predictions for 2011 using 2010 as a reference year, for example, would have been slightly lower (that is, harsher) at grade C using the alternative measures such as K2RwG10EM, whereas they would have been more lenient for 2013 (using either 2011 or 2012 as a reference year). This may be explained by the particular specifications being compared in a time of specification change.

For 2012-2013, at all three grades shown, most measures had median differences slightly above zero, indicating that predictions would be higher (more lenient) using the alternative methods in most subjects. In particular, there appears to be a systematic difference between the methods that rely upon an explicit grade inflation adjustment (the five boxes at the left of each plot) and those that are based upon normalised scores or quantiles (where the grade inflation adjustment is implicit). On inspection, this is particularly acute in high-performing subjects such as Biology, Chemistry, Physics and French. This issue is explored further in Section 2.7.1.

As might be expected, the measure that results in predictions closest to the current method, and with the smallest interquartile range, is K2LevGEM, using mean level in English and Mathematics only. The differences for quantiles based on English and Mathematics raw scores only (K2RwG8EM, K2RwG10EM) are markedly different from those based on English, Mathematics and Science (K2RwG8, K2RwG10). This is caused by the difficulty of breaking raw scores into precise quantiles given that groups are defined by a limited number of whole marks that pupils can achieve at KS2. For example, for KS2 candidates in 2007 (those who were in Year 11 in 2012) 13.6 per cent of candidates were assigned to the 5th octile by raw scores including Science compared with 12.2 per cent from raw scores excluding Science²⁵. However, for KS2 candidates in 2008 (those who were in Year 11 in 2013) 13.0 per cent of candidates were assigned to the 5th octile by raw scores including Science compared with 13.3 per cent from raw scores excluding Science. In other words, while the percentage of candidates in a particular octile decreases between years for K2RwG8, it increases for K2RwG8EM. In summary, the distribution of prior attainment changes in slightly different ways from the reference to the outcome year depending on the measure used to construct the quantiles.

²⁵ Ideally all octiles should contain exactly 12.5 per cent of pupils nationally.

Figure 2.12: Differences of predictions from current method at grade C (cumulative)

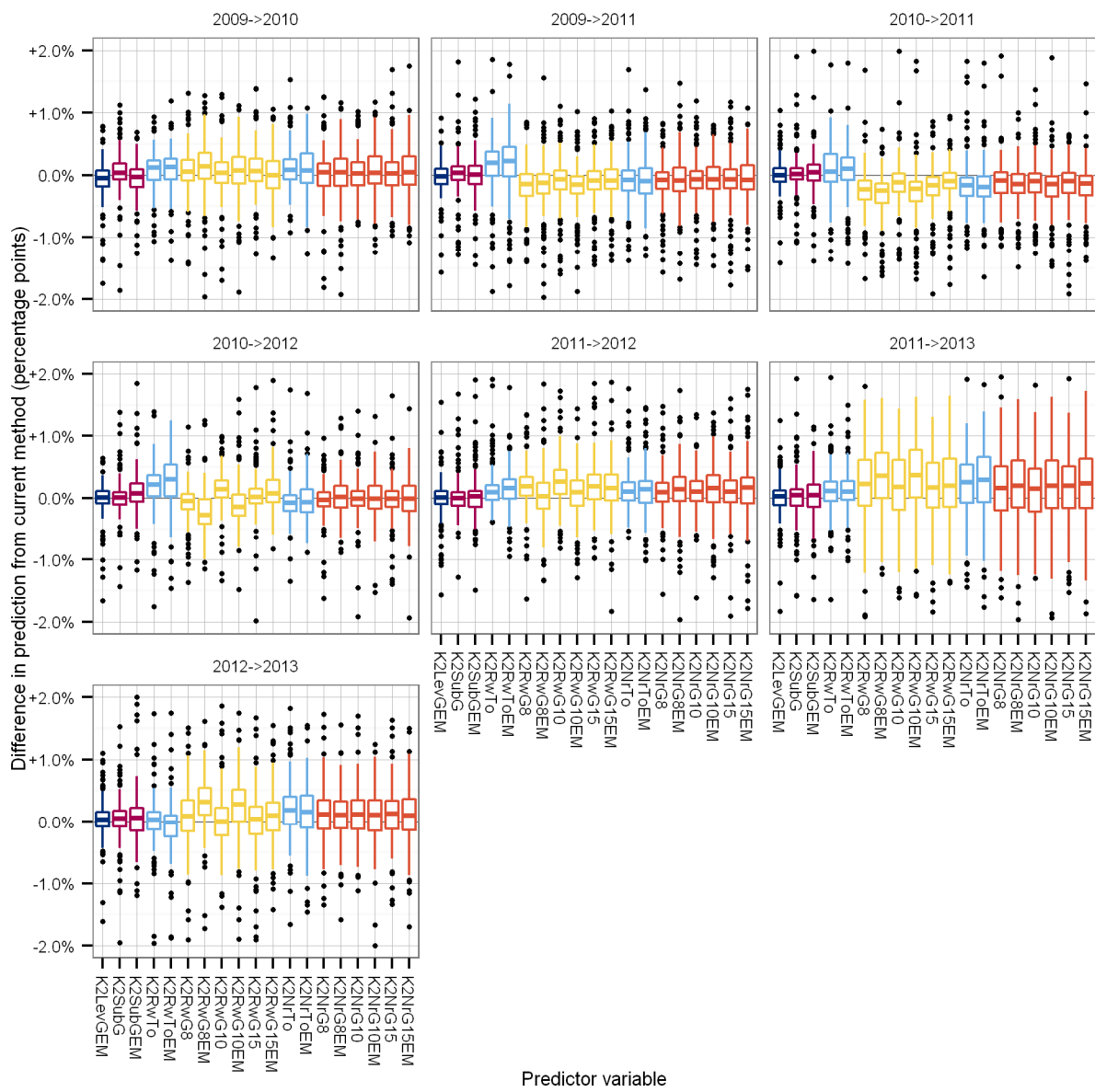


Figure 2.13: Differences of predictions from current method at grade F (cumulative)

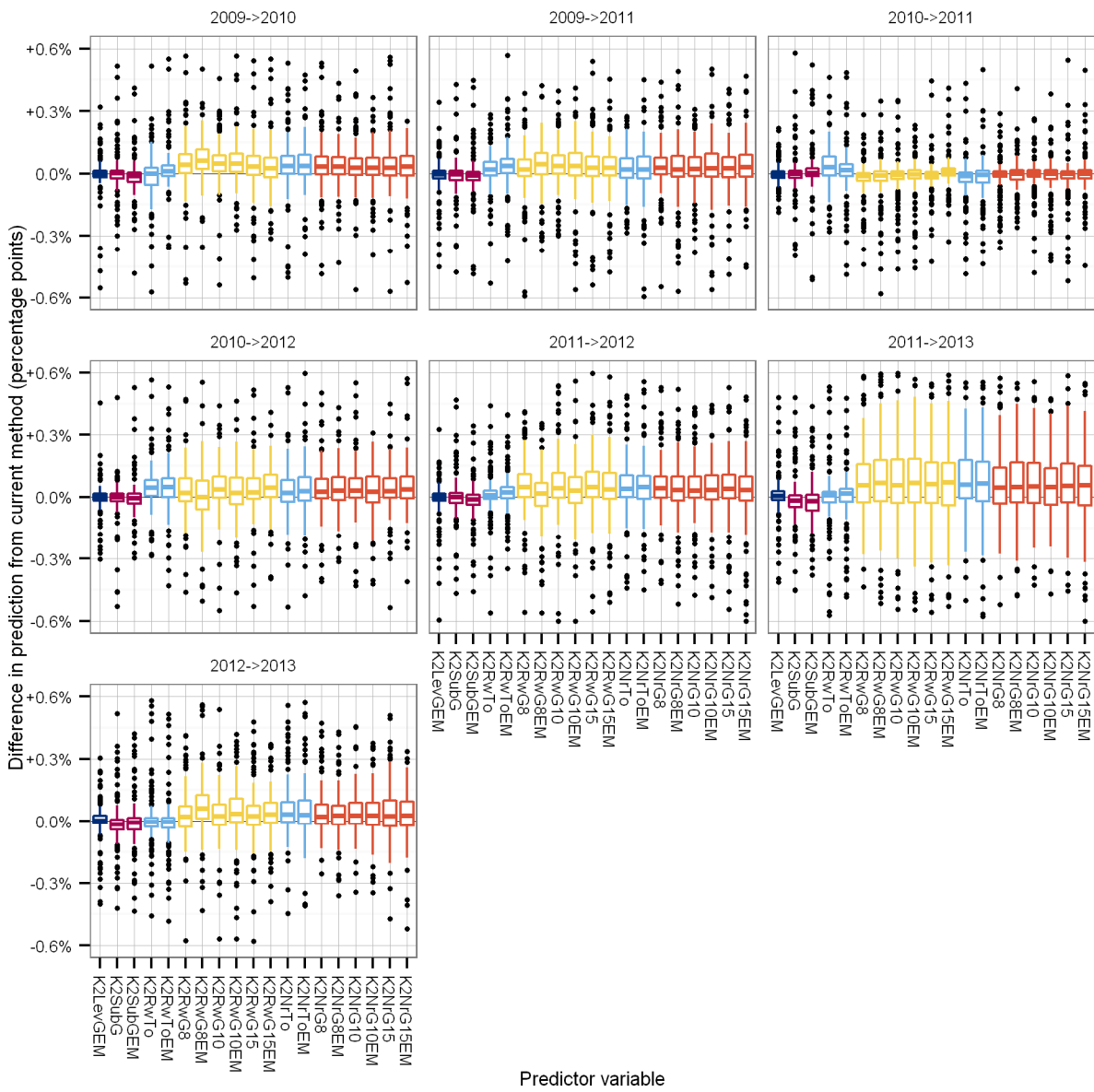


Figure 2.14: Differences of predictions from current method at grade A (cumulative)

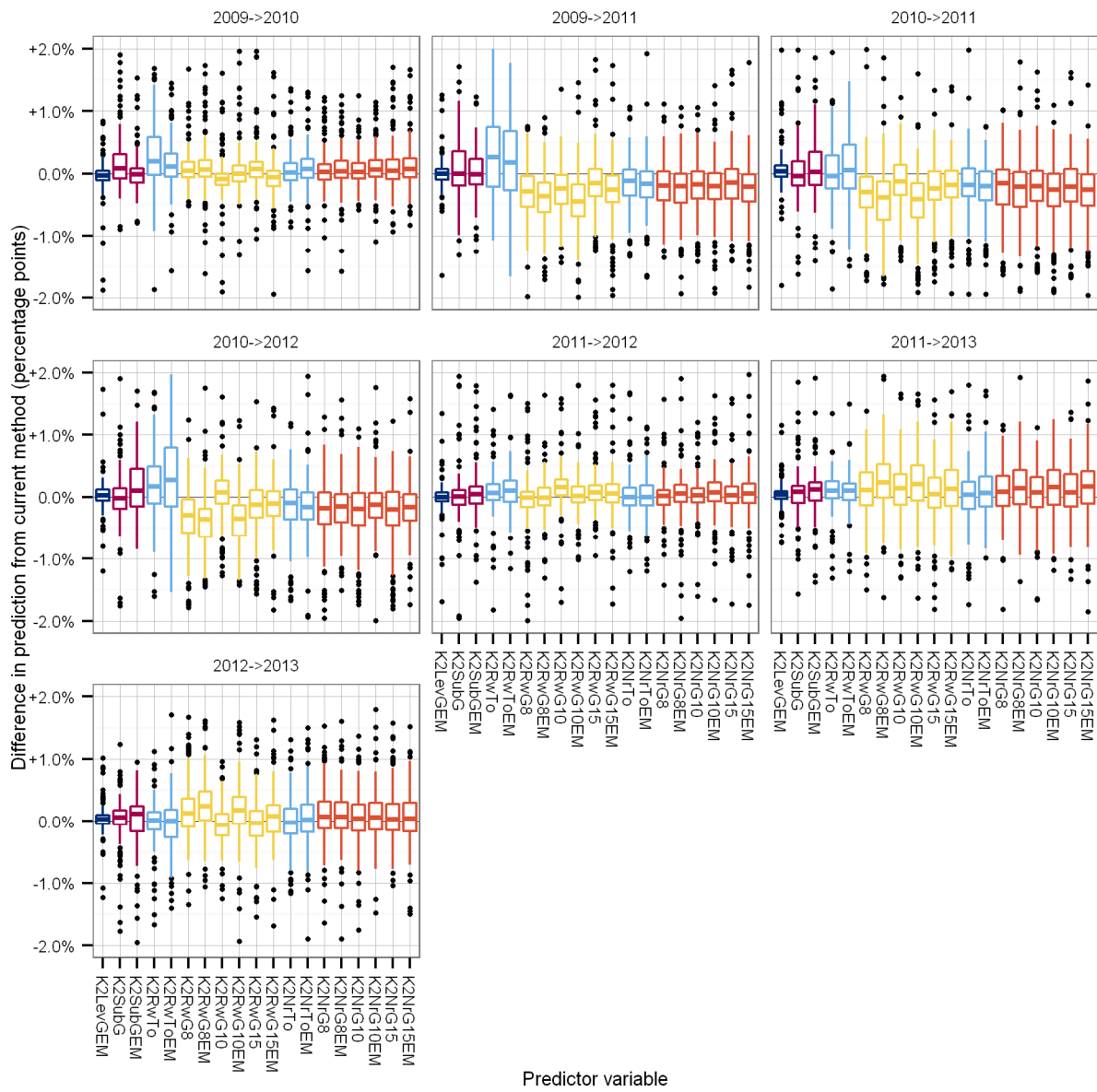


Table 2.9: Differences of predictions compared to current method (percentage points)

Predictor variable	A		C		F	
	Median	IQR ²⁶	Median	IQR	Median	IQR
K2LevGEM	0.023	0.126	0.019	0.203	0.002	0.034
K2SubG	0.076	0.212	0.044	0.233	-0.017	0.044
K2SubGEM	0.105	0.340	0.051	0.324	-0.012	0.047
K2RwTo	0.024	0.260	0.017	0.269	-0.007	0.043
K2RwToEM	-0.006	0.385	-0.020	0.311	-0.005	0.044
K2RwG8	0.133	0.438	0.038	0.454	0.019	0.092
K2RwG8EM	0.231	0.415	0.319	0.409	0.064	0.108
K2RwG10	-0.067	0.348	-0.002	0.415	0.020	0.087
K2RwG10EM	0.173	0.406	0.270	0.428	0.038	0.106
K2RwG15	-0.014	0.362	0.036	0.408	0.021	0.091
K2RwG15EM	0.076	0.413	0.073	0.431	0.029	0.087
K2NrTo	-0.024	0.383	0.166	0.415	0.031	0.095
K2NrToEM	0.024	0.405	0.149	0.464	0.028	0.108
K2NrG8	0.088	0.399	0.080	0.430	0.021	0.089
K2NrG8EM	0.069	0.385	0.087	0.402	0.027	0.091
K2NrG10	0.072	0.396	0.105	0.406	0.026	0.092
K2NrG10EM	0.056	0.374	0.072	0.457	0.023	0.090
K2NrG15	0.053	0.396	0.089	0.443	0.020	0.102
K2NrG15EM	0.071	0.439	0.110	0.450	0.025	0.102

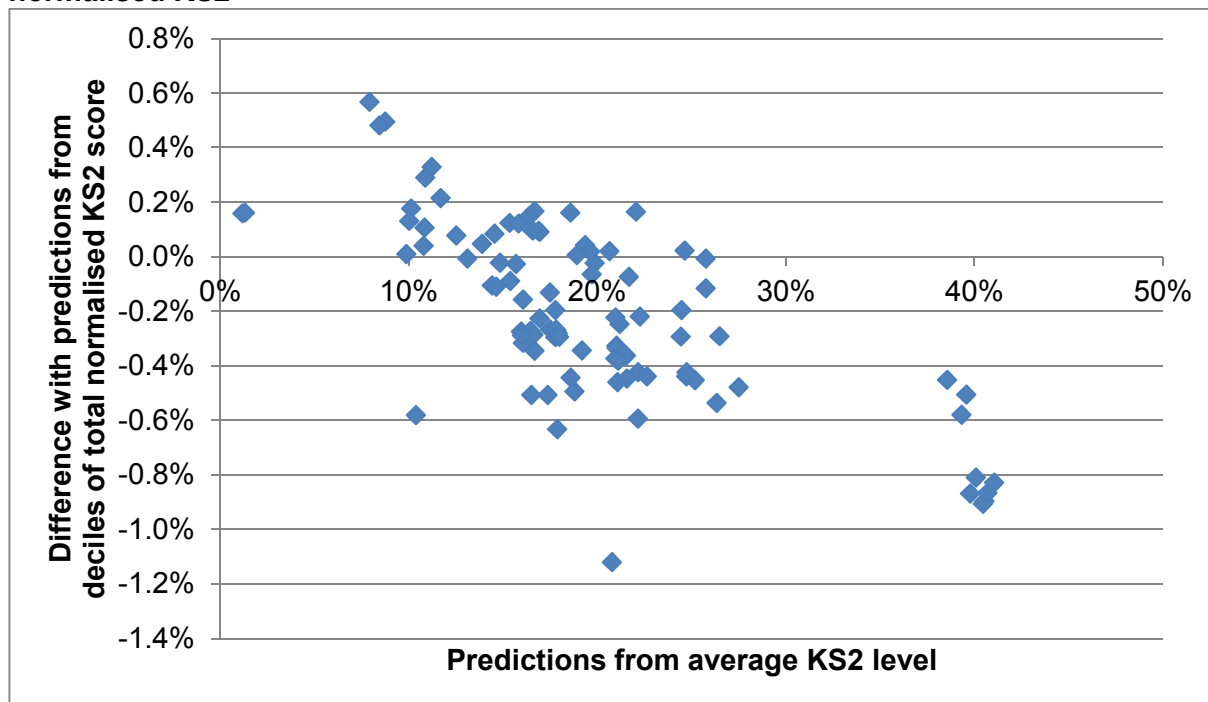
2.7.1 Further exploration of the effect of the KS2 grade inflation adjustment

As has been demonstrated above, our analysis suggests that in certain years for the highest performing subjects, results based upon KS2 levels may be systematically different from those predicted using a method that is not reliant upon the grade inflation adjustment. For example, a method based upon quantifying KS2 attainment in terms of normalised scores or deciles. This section explores this phenomenon further.

The effect of interest is displayed in Figure 2.15. It should be noted that because the size of this effect is relatively small (usually associated with less than 1 percentage point of difference between predictions) this analysis is restricted to subjects with more than 3,000 candidates. In other words, figure 2.15 only includes those awards where the current recommended tolerance for differences between actual and final outcomes is 1 percentage point. The results focus upon differences at grade A. Because the largest consistent differences occur for the Single Science subjects, and because these subjects were awarded for 2013 based upon a reference year of 2011, Figure 2.15 is based upon these years.

²⁶ Inter-quartile range.

Figure 2.15: Differences between predictions from KS2 average level and deciles of total normalised KS2



As illustrated in Figure 2.15, there is a very clear negative association between the percentage of candidates predicted to achieve grade A or above from average KS2 level, and the difference with predictions using deciles of normalised scores. That is, the predictions from KS2 levels are too low for subjects with high ability candidates whereas they tend to be slightly too high for subjects with lower ability candidates. Whilst these differences are fairly small, given the tight tolerances applied to subjects with entries of this size, they could have a noticeable impact. In particular, towards the right hand side of Figure 2.15, are the predictions for the Separate Sciences for each AO and it is evident that these predictions are consistently between 0.4 and 1.0 percentage points lower than would have been predicted using a method not dependent upon the explicit KS2 grade inflation adjustment²⁷.

The reason for these differences is contained within the way the KS2 grade inflation adjustment is applied to each subject. At present, the grade inflation adjustment works by calculating a predicted grade distribution for each subject if all KS2 candidates nationally were to enter it. This is done using the prediction matrix derived in the reference year for the national KS2 distributions five years prior to both the reference year and the outcome year²⁸. The difference between these two sets of predictions is then used to adjust the predictions in the outcome year (see Appendix 1).

The weakness with the above technique is that it is applied to each subject (and each AO within that subject) as a blanket adjustment with no regard for the differences in the prior attainment distributions of the candidates to which it is being applied. The weakness in this approach is explored further below.

To begin with let us compare the distribution of KS2 attainment, in terms of average levels, between the national populations in 2006 and 2008. That is, the populations associated with taking GCSEs in 2011 and 2013. A comparison of the two distributions is shown in Table 2.10.

²⁷ Similar results to those displayed in Figure 2.15 can be obtained by comparison with KS2 groupings of 8, 10 and 15 groups based on KS2 total raw scores, KS2 total raw scores (including or excluding Science) and also by comparisons with considering normalised scores as continuous predictors and using logistic regression.

²⁸ That is, the year in which we are interested in setting grade boundaries.

The table shows that in the national population a smaller percentage of candidates are in the top KS2 categories in 2008 than in 2006. For example, whereas in 2006 21.1 per cent of candidates achieved an average level of 5 or above, only 18.5 per cent achieved this in 2008. Furthermore, whereas in 2006 36.9 per cent of candidates achieved an average level of 4.66 or above, only 34.9 per cent achieved this level or above in 2008. This could potentially be interpreted as implying that KS2 became more difficult towards the upper end of the scale between 2006 and 2008. In contrast, at the lower end of the ability distribution there is evidence of increased attainment at KS2. For example, whereas in 2006 72.2 per cent of candidates achieved at level 4 or above, by 2008 this percentage had risen to 74.8. Already, this implies that, a blanket grade inflation adjustment may be inappropriate because KS2 may have become more difficult at some points on the scale and easier at others. Thus, any adjustment for “grade inflation” at KS2 may need to account for differences in the distribution of candidates across different levels.

Table 2.10: A comparison of the cumulative national distributions of KS2 levels in 2006 and 2008 and the associated probability of achieving grade A or above in Biology in 2011

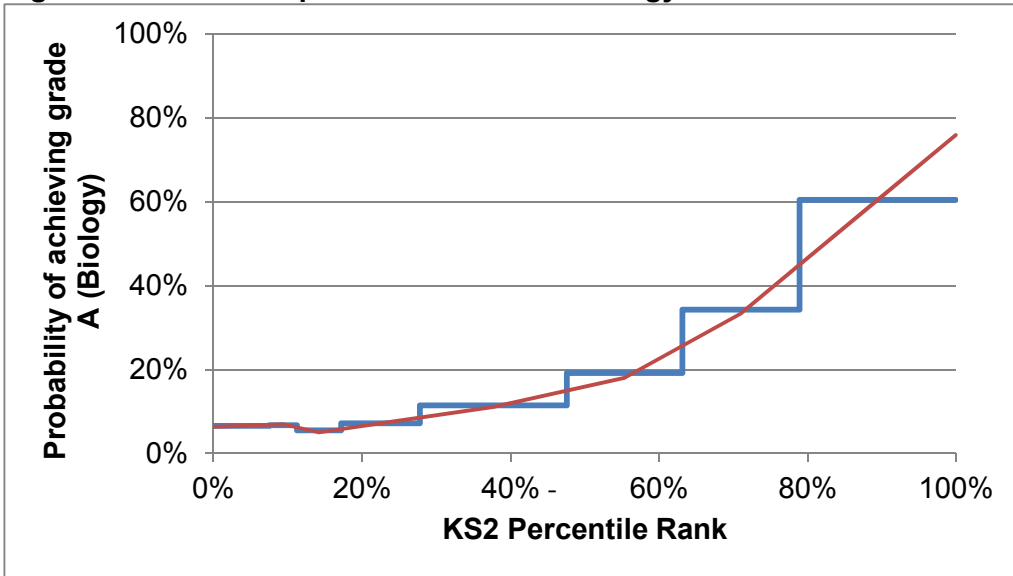
Average KS2 level	Percentage of KS2 population in each group or above		Probability of achieving grade A or above in GCSE Biology (2011)
	2006	2008	
<3.00	100.0%	100.0%	6.6%
3.00	92.5%	93.4%	6.8%
3.33	88.8%	90.3%	5.5%
3.66	82.8%	84.9%	7.2%
4.00	72.2%	74.8%	11.4%
4.33	52.4%	51.8%	19.1%
4.66	36.9%	34.9%	34.3%
5.00	21.1%	18.5%	60.4%

The final column of Table 2.10 also shows the probability of candidates achieving an A or above in GCSE Biology in 2011 dependent upon their level of prior attainment in 2006. As can be seen, higher levels of prior attainment are (usually²⁹) associated with an increased chance of achieving an A in Biology GCSE. For example, only 6.6 per cent of candidates averaging below level 3 at KS2 in 2006 achieved a grade A or above in Biology GCSE in 2011. In contrast, 60.4 per cent of those candidates with an average KS2 level of 5 achieved a grade A or above in Biology GCSE in 2011.

This prediction matrix can be imagined visually as shown by the blue line in figure 2.16. The x-axis converts each of the KS2 categories in Table 2.10 into percentile ranks and then the y-axis shows the probability of achieving a grade A or above associated with each grouping. For example, Table 2.10 shows that in 2006 the top 21.1 per cent of KS2 candidates achieved an average level of 5 in Biology and that 60.4 per cent of these candidates (of those that also took Biology GCSE) achieved a grade A or above. Thus the blue line in Figure 2.16 between the percentile ranks of 78.9 and 100 is at a probability of 60.4 per cent. The slightly crude representation of the link between KS2 percentile ranks and achieving a grade A displayed by the blue line is converted into a more plausible continuous relationship by the red line. The red line is devised so that its average height within any region defined by the position of the “steps” on the blue line is equal to the probability defined by the blue line. For example, between the percentile ranks of 78.9 and 100 the red line is designed to have an average height of 60.4 per cent.

²⁹ The slight drop in probability seen for those candidates with an average KS2 level of 3.33 is based on roughly 500 candidates only and so is likely to be caused simply by random variation.

Figure 2.16: Visual representation of the Biology Prediction Matrix

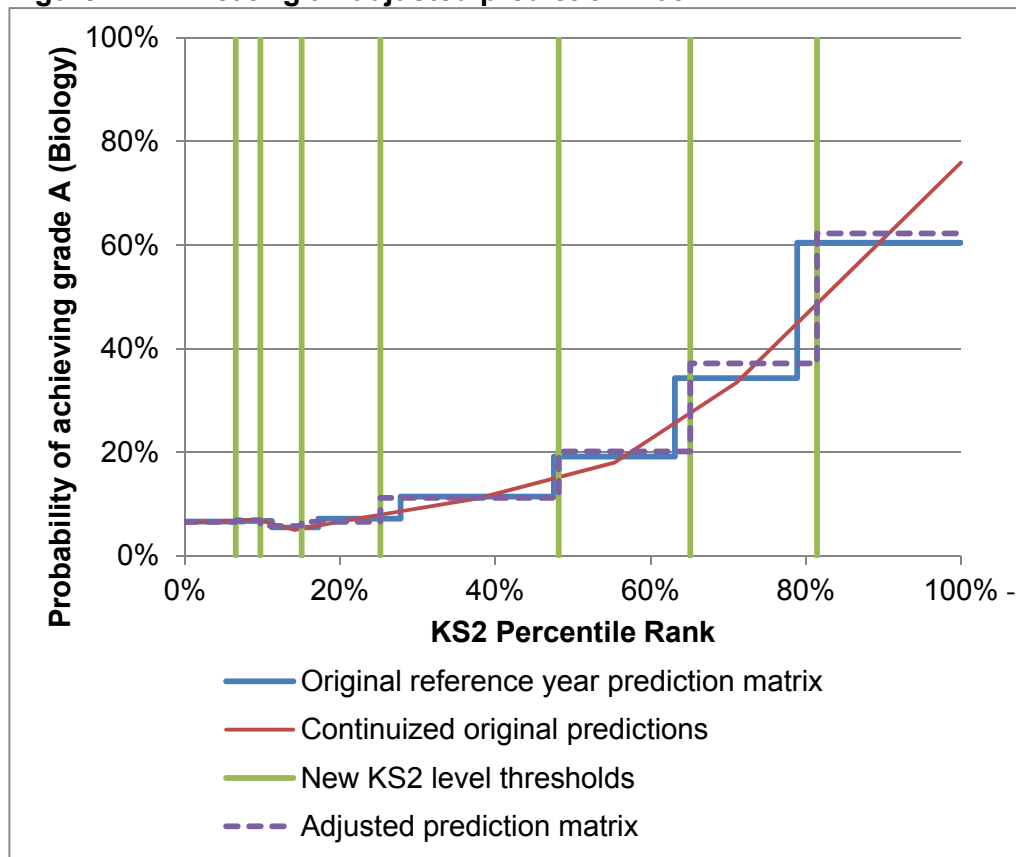


Having established a plausible continuous relationship between KS2 percentile ranks and achievement at GCSE (the red line in Figure 2.16), we can now use this relationship to generate an expected relationship between average 2008 KS2 levels and the probability of achieving grade A or above in Biology GCSE. That is, the expected 2013 prediction matrix. This process is shown in figure 2.17. The green lines represent the “percentile boundaries” between the KS2 groupings in 2008. For example only 18.5 per cent of candidates were in the top KS2 category in 2008, therefore a green line is positioned at 81.5 on the x-axis. Having established these “percentile boundaries” between KS2 categories, we can then calculate the average value of the red line within each category (the purple line). For example, within the region of figure 2.17 defined by the top category, the average height of the red line is 62.3 per cent. If we assume that the relationship between KS2 achievement percentiles and GCSE attainment is constant over time³⁰, then this implies that a candidate with an average KS2 level of 5 in 2008 has a 62.3 per cent chance of achieving grade A or above in 2013.

³⁰ An assumption which is more or less fundamentally at the heart of applying a KS2 grade inflation adjustment at all.

Examining Figure 2.17 more closely we see that the prediction matrix for grade A should change substantially for the top two KS2 categories, but hardly at all for the other categories. For example, in the top KS2 category the probability of achieving grade A or above is adjusted from the original 60.4 per cent to 62.3 per cent. In the next category down the probability is adjusted from 34.3 per cent to 37.2 per cent. In contrast, two categories below this, the probability is hardly adjusted at all; from 11.4 per cent to 11.2 per cent. The calculations imply that the adjustment that should be applied to an AO should depend upon the prior attainment distribution of their candidates. The more candidates they have in the top two prior attainment categories³¹, the higher the (positive) grade-inflation adjustment should be. Furthermore, for subjects such as Separate Sciences, where the prior attainment distribution of candidates is especially high relative to the national KS2 distribution, a grade inflation adjustment based on extrapolated national predictions will underestimate the size of the adjustment that is necessary. In fact, because high ability candidates tend to take greater numbers of GCSEs, the grade inflation adjustment will have been slightly underestimated for the majority of subjects. This is why in Figure 2.15 the predictions from deciles of normalised KS2 scores are very slightly higher than those from average KS2 levels for the majority of subjects. This may explain the fact noted in Section 2.7 that for many subjects predictions based upon alternative measure of KS2 were very slightly higher than those based upon KS2 levels.

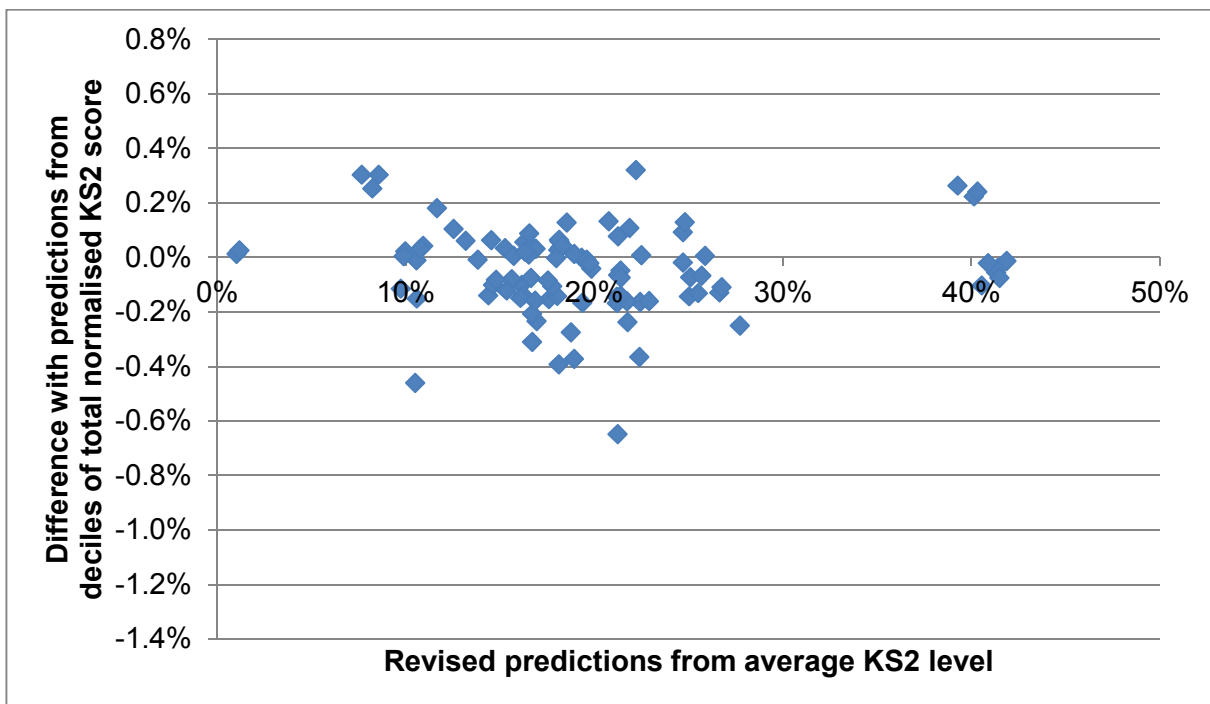
Figure 2.17: Creating an adjusted prediction matrix



³¹ Particularly the second from top category.

The method of adjusting the prediction matrix as described here (as opposed to applying a blanket KS2 grade inflation adjustment) was practically applied to the data for each subject/AO combination shown earlier in Figure 2.15. The differences between these predictions and those derived from deciles of normalised KS2 scores are shown in Figure 2.18. As can be seen there is no longer a tendency for the predictions based on KS2 levels to be lower than the decile-based predictions for subjects with high ability candidates. This confirms that our diagnosis of the reason for the original differences is correct.

Figure 2.18: Differences between revised predictions from KS2 average level and deciles of total normalised KS2



It is worth noting that the impact of the existing grade inflation adjustment technique will not necessarily lead to GCSE grade boundaries being set more harshly than a decile-based method every year. The direction and size of the difference will depend upon the national KS2 distribution each year. Table 2.11 shows the national distribution of average KS2 levels over time. This shows that, whereas the proportion of candidates in the higher KS2 categories decreased from 2006 to 2008, there was an increase from 2004 to 2006. This means that, in contrast to the situations explored above, GCSE grade boundaries set using average KS2 levels in 2011 would be likely to be too generous for high attaining GCSE subjects such as Separate Sciences. Looking forwards to 2014 GCSEs, the KS2 grade inflation adjustment may again lead to overly harsh boundaries for such subjects if based on 2012 as the reference year. However, because the national KS2 distribution was fairly consistent between 2008 and 2009, if predictions for 2014 are based on 2013 as the reference year, then the grade inflation adjustment is likely to have very little impact at all as it will be close to zero for all subjects.

Table 2.11: A comparison of the cumulative national distributions of KS2 levels over time

Average KS2 Level	Percentage of KS2 population in each group or above in each year (year in which candidates took/will take GCSEs – aged 16 – in parentheses)					
	2004 (2009)	2005 (2010)	2006 (2011)	2007 (2012)	2008 (2013)	2009 (2014)
<3.00	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
3.00	91.9%	92.4%	92.5%	93.0%	93.4%	93.7%
3.33	88.2%	88.7%	88.8%	89.6%	90.3%	90.4%
3.66	81.6%	82.5%	82.8%	84.1%	84.9%	84.7%
4.00	70.3%	71.9%	72.2%	73.6%	74.8%	74.1%
4.33	49.3%	51.7%	52.4%	53.2%	51.8%	51.7%
4.66	33.3%	34.4%	36.9%	37.3%	34.9%	35.2%
5.00	17.6%	17.4%	21.1%	21.3%	18.5%	18.6%

Taking the discussion above as a whole, it is clear that there are three possible ways to address the current weakness in the KS2 grade inflation adjustment:

1. - Use measures of KS2 that do not require such adjustments. This could include any method based upon normalised scores or any of the methods based upon groupings (such as deciles) using total raw scores at KS2. Earlier analysis has already shown that many of these measures are very slightly better predictors of achievement at KS2 in any case. For this reason, this would be our recommended approach.
2. - Continue using KS2 levels but apply a more nuanced grade inflation adjustment that takes account of the different prior attainment distributions for different subjects and different AOs. One possible technique to do this has been described above.
3. - Stop applying any kind of KS2 grade inflation adjustment. This would require greater faith that the changes in the KS2 level distribution over time reflect genuine changes in the ability of the national cohort. To some extent this position is justifiable due to the relatively strong nature of the mechanisms (including the use of pre-testing and anchor tests) used to ensure standards are maintained at KS2. However, such an approach would also require an acceptance that GCSE pass rates would fluctuate year on year in line with patterns established for the different cohorts at KS2. In order to minimise such fluctuations it may be desirable to choose the reference years used to generate GCSE predictions so that there is as little change over time as possible. For example, this might include specifying 2013 as the most appropriate reference year to set GCSE standards in 2014.

2.8 Summary

The analysis in this section has shown that:

- Predictions based on KS2 levels are very similar to those based on more detailed ways of quantifying KS2 attainment such as sub-levels, raw scores and normalised scores.
- Some small gains in predictive power could be achieved by using an alternative measure and, of the ones considered, the best measure would be to use logistic regression based upon total normalised scores.
- The loss of KS2 Science is likely to have only a minor impact on predictive power of models or the values of the predictions themselves³².
- KS2-based predictions of how far each AO's candidates should be from the national level of attainment in each GCSE subject are highly correlated with predictions based upon concurrent attainment.
- Having said this, compared to predictions based on concurrent attainment, KS2-based predictions tend to under-predict the likely extent of differences between AOs. This issue will be explored further in Section 4.
- The current KS2 grade inflation adjustment suffers from not taking account of differences in the prior attainment distribution of candidates in different subjects and in different AOs. This weakness could be addressed through amendments to the calculations but would be automatically addressed if calculations were based on logistic regression using normalised scores or another method not dependent upon the comparability of KS2 levels over time.

³² Further exploration of the data, looking specifically at GCSE Science subjects, found that in 2013, the removal of Science KS2 did not lead to a change in predictions (at any grade) of more than 0.5 percentage points for any of these subjects.

3. Review of tolerances for reporting outcomes that do not meet predictions

Every summer Ofqual publishes data exchange procedures for GCE and GCSE certificates. This regulatory document states, amongst other things, the *reporting tolerance*³³ for GCSEs. Specifically the document requires that

“Wherever actual and predicted outcomes differ for grades A and C beyond a given reporting tolerance, depending on entry size, the relevant AO will inform the regulators and other AOs of the details.” (Ofqual 2013³⁴, page 6).

In this context, informing the regulator “of the details” means that each AO must provide evidence justifying why the difference from predicted outcomes is necessary. At present the reporting tolerances for GCSE are as follows:

Table 3.1: Current GCSE reporting tolerances

Number of candidates with matching KS2 data available	Reporting tolerance
Less than 500	No reporting tolerance is applied
501-1000	3 percentage points
1001-3000	2 percentage points
3001+	1 percentage point

Given that AOs are required to justify differences between predicted and actual outcomes that are bigger than tolerance, it is clearly desirable that differences of this nature are unlikely to occur purely by chance. For example, if an out of tolerance difference between predicted and actual outcomes could be caused simply by the ordinary variation in achievement between different schools then it may prove difficult for an AO to provide further evidence (beyond the expert opinion of examiners) justifying this.

The current tolerances noted above are based upon research into AS and A level predictions using prior attainment at GCSE (Benton and Lin, 2011). There are some reasons to suspect that the tolerances calculated in this scenario do not directly apply to GCSE predictions using KS2. Firstly, the correlation between KS2 and GCSE is somewhat lower than the correlation between (mean) GCSE and AS/A level, leading to the possibility that slightly wider tolerances may be required in the former case. Furthermore, the number of entries per centre is likely to be larger at GCSE than at A level. This means that an entry of a given number of candidates at GCSE is likely to come from a smaller number of centres than an entry of the same size at A level. This again may imply that, for any fixed number of entries, the tolerance at GCSE should be wider than the tolerance at A level³⁵.

The aim of the analysis presented in this section is to derive new tolerances for GCSE based upon specific analysis of achievement at GCSE.

³³ The word “tolerance” itself may be unhelpful as it carries connotations of quality control within a manufacturing process where provided each component is constructed to within a given tolerance level we can be certain the system as a whole will function. As will be explored further, this is not the same as the thinking underpinning the tolerances used for GCSE awarding. A more correct term might be *justified variation from expected outcomes*. However, in order for consistency with existing documentation, the term tolerance will be retained within this report.

³⁴ <http://ofqual.gov.uk/files/2013-06-06-summer-2013-data-exchange-procedures.pdf>

³⁵ Although, of course, for most subjects the number of candidate entries at GCSE far exceeds the number at A level.

3.1 Method and results

The method used to derive tolerances is essentially the same as that used in the previous research into tolerances at AS and A level; namely *balanced repeated replication* (BRR). The basic idea of BRR is to repeatedly recalculate the quantities we are interested in based upon a randomly chosen half of the available centres within the data. The extent of variation between different half samples is related to the standard error of the quantity we are interested in by a known formula. In other words if we get very different answers when we recalculate the quantity of interest with different half samples we know that there is a large standard error. If different half samples give very similar results then the standard error must be small. For the analysis in this section, the quantity we are interested in is the standard error of the difference between predicted and actual results. In broad terms, for each AO and each GCSE subject, we have calculated tolerances by comparing the difference between actual and predicted results using one half-sample of centres to the difference between actual and predicted results using another half-sample of centres. If the difference varies wildly, for example actual achievement being considerably above predictions in one half-sample and considerably below predictions in another, then we know that the method requires a large tolerance. If, however, the differences are very consistent between different half-samples³⁶ then we know smaller tolerances may be sufficient. Full details of the procedure used to calculate tolerances are given in Appendix 2.

In line with current practice tolerances were estimated to represent 75 per cent confidence intervals³⁷. This means that if we had an independent means of knowing the “correct” grade boundary for each subject for each AO, and further that the model underlying prediction matrices was true, then the correctly awarded outcomes for any subject within any AO would be within tolerance of predictions three quarters of the time.

Tolerances were estimated for each GCSE subject awarded by each AO in June 2013. Estimates were based on predictions generated using the performance of 16 year old candidates in both June 2011 and June 2012. These tolerances are plotted against the number of matched candidates taking each subject with the AO in Figures 3.1. Smooth lines showing how the average level of estimated tolerance at grades A and C changes dependent upon the number of candidates are included within each chart. As discussed elsewhere (Benton and Lin 2011, Smith 2013), the reliability of predictions will depend not only upon the number of candidates in the outcome year but also on the number of candidates used to construct the prediction matrix and the proximity of the prediction to 50 per cent³⁸. However, given that current guidelines focus upon the number of candidates for whom predictions have been made, we have chosen to focus upon this factor as of primary importance here. In order to better display the general trend one outlying subject (AQA Environmental Science) has been excluded from this chart. This subject displayed an unusually high estimated tolerance of more than 11 percentage points – probably caused by the fact that the majority of candidates for this subject were located in a very small number of centres. Subject/AO combinations with more than 20,000 candidates have also been excluded from this graph. Such subjects universally had estimated tolerances below 1.4 percentage points with all but four estimated tolerances³⁹ below 1 percentage point. Although there was some evidence of tolerances continuing to decrease beyond 20,000 candidates the scale of the decreases were small and are difficult to display visually on the same chart as the smaller awards.

³⁶ Note that we do not require actual and predicted outcomes to match. Only that the difference, which may be as the result as leniency or severity on the part of an AO, is consistent between half-samples of centres.

³⁷ This is the basis of tolerances in the report into AS and A levels (Benton and Lin 2011). Furthermore, once Science subjects (where additional adjustments were openly made to predictions) are excluded from analysis, we found that (using our predictions) 36 out of 130 awards (28%) were out of tolerance at grade C whereas 23 out of 130 (18%) were out of tolerance at grade A. Thus, approximately 75% of awards were within tolerance in practice.

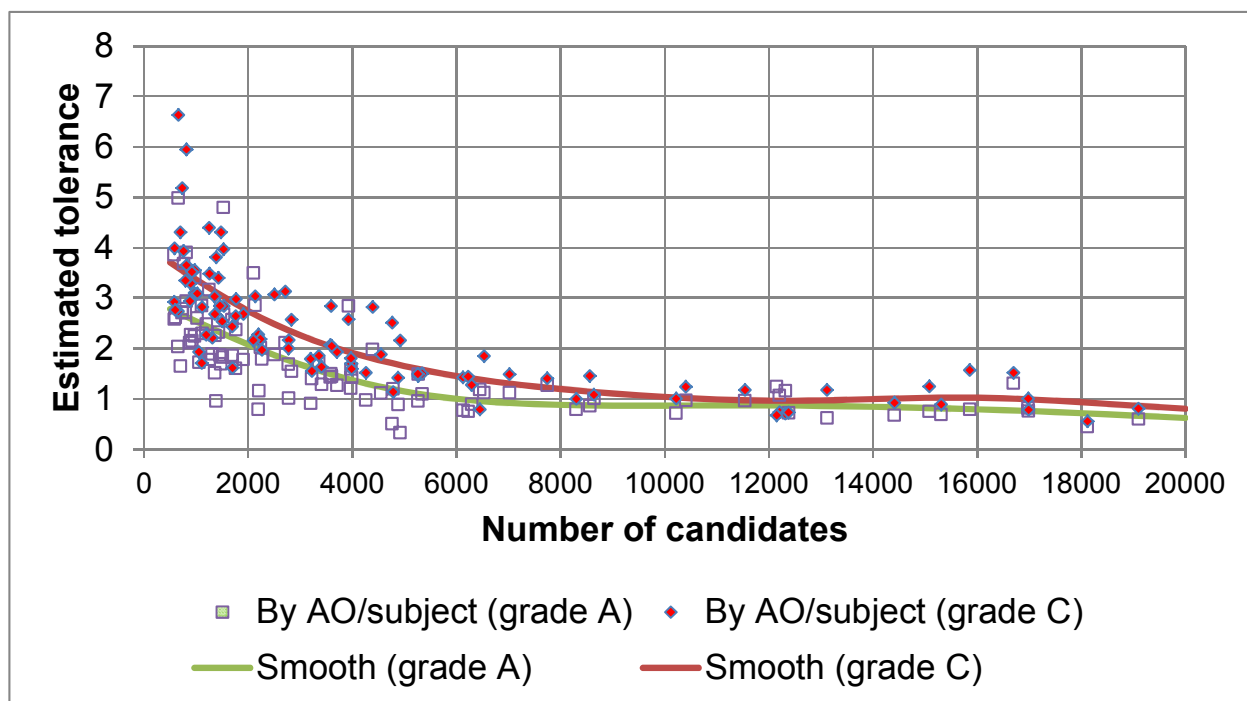
³⁸ Probabilities close to 50% are subject to greater fluctuation between different samples than probabilities close to either 0% or 100%.

³⁹ Out of a total of 48 subject/AO combinations with more than 20,000 candidates included in analysis, only 1 tolerance larger than 1 percentage point was found for grade A with 3 found for grade C.

In general Figure 3.1 shows a strong relationship between the number of candidates entering a qualification and the estimated tolerance. At grade A the estimates are roughly in line with the current recommended tolerances. For example, the average level of tolerance appears to fall below 2.5 percentage points at roughly 1,000 candidates indicating that a tolerance level of 2 percentage points would be more appropriate than a tolerance level of 3 percentage points for this number of candidates. Similarly, the average level of tolerance falls below 1.5 percentage points at just above 3000 candidates indicating a guideline tolerance of 1 will be more appropriate than 2 for this number of candidates.

Estimated tolerances at grade C tend to be higher than those for grade A. Indeed, for the 147 subject/AO combinations analysed the estimated C grade tolerance was higher than the estimated A grade tolerance on 123 occasions (84 per cent). Furthermore, the analysis shows that roughly 2500 candidates are required for the average tolerance to fall below 2.5 percentage points and around 6000 are required for it to fall below 1.5 percentage points. This indicates that the guidelines should be amended to allow for greater tolerances at grade C than at grade A.

Figure 3.1: Relationship between estimated tolerances and number of candidates



Another potential weakness of the current guidelines is the sudden step changes in the sizes of tolerances according to the number of candidates. For example, an award based on 1000 matched candidates has a tolerance of 3 percentage points whereas if there are 1001 matched candidates the tolerance drops all the way to 2. The effect of these step changes is explored further in Figure 3.2.

Based on the average tolerance levels displayed by the smooth lines in Figure 3.1⁴⁰, Figure 3.2 shows the estimated probability that the correctly awarded outcomes for a subject/AO combinations would be outside of current tolerance guidelines⁴¹. This shows that, even at grade A, where we have seen that current tolerances are roughly in line with analysis, the probability of an award being out of tolerance varies considerably depending on the number of candidates. If the number of candidates is just above a particular threshold then the probability of an award being out of tolerance can be considerably larger than the target of 25 per cent. At worst, for

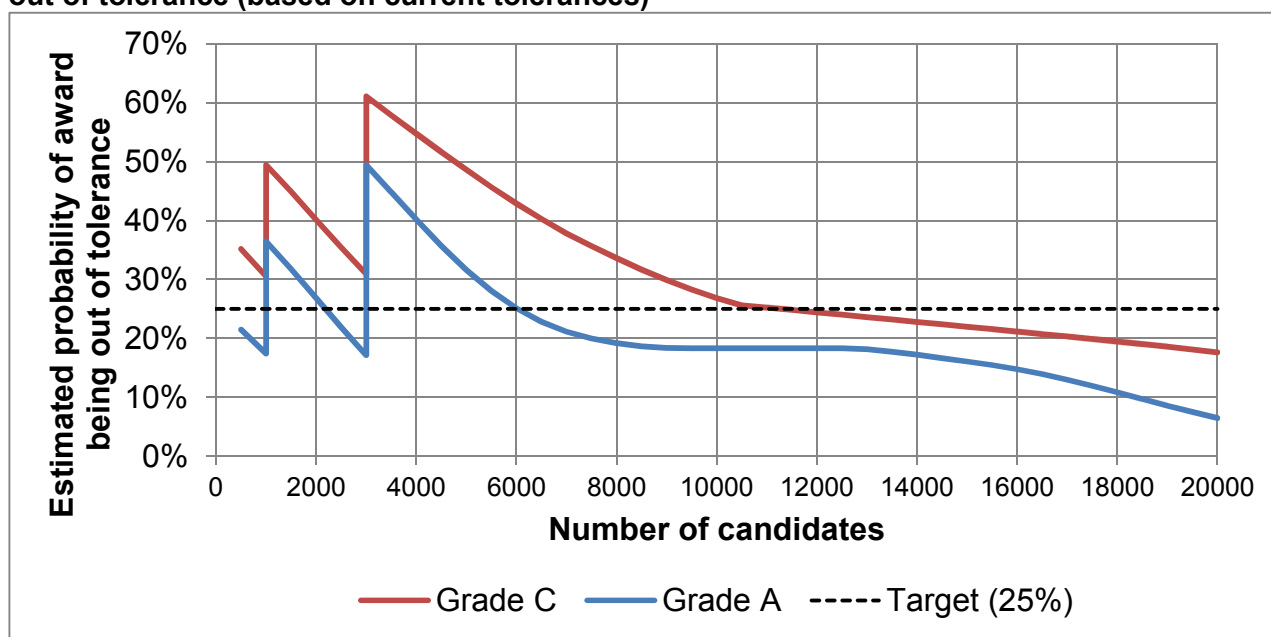
⁴⁰ But marginally adjusted to ensure that the tolerances are monotonically decreasing as the sample size increases.

⁴¹ Calculated using a simple normal approximation formula.

awards based on 3001 candidates there is almost a 50 per cent chance of correctly awarded outcomes being judged out of tolerance. On the other hand, if the number of candidates is far greater than the threshold, the probability of a correctly awarded GCSE being out of tolerance with predictions can be considerably below 25 per cent.

At grade C, Figure 3.2 shows that unless there are at least 10,000 matched candidates, the probability of an award being out of tolerance universally exceeds 25 per cent. This is partially caused by the fact that, as described earlier, estimated tolerances at grade C tend to be higher than at grade A. However, this effect is exacerbated by the step changes in the currently recommended tolerances. At worst, for awards based on 3001 candidates there is a probability of more than 60 per cent that a correctly awarded GCSE would be out of tolerance with predictions.

Figure 3.2: Estimated probabilities of correctly awarded GCSEs having outcomes that are out of tolerance (based on current tolerances)



Given the findings above, we would strongly recommend that the current guidelines are amended to allow for different, higher tolerances at grade C. If possible the recommended tolerances should also be more finely grained to reduce the effect of step changes on the chances of corrected awarded GCSEs being judged out of tolerance. A set of possible alternative guidelines is detailed in Table 3.2. An alternative simple method to derive tolerances is provided in the following section.

Table 3.2: Recommended revised tolerances for GCSE awarding

Recommended tolerance	Sample size for tolerance to be applicable at each grade	
	Grade C	Grade A
3%	500-2000	NA
2.5%	2001-3000	500-1500
2%	3001-4500	1501-2500
1.5%	4500-7500	2501-4500
1%	7501+	4500+

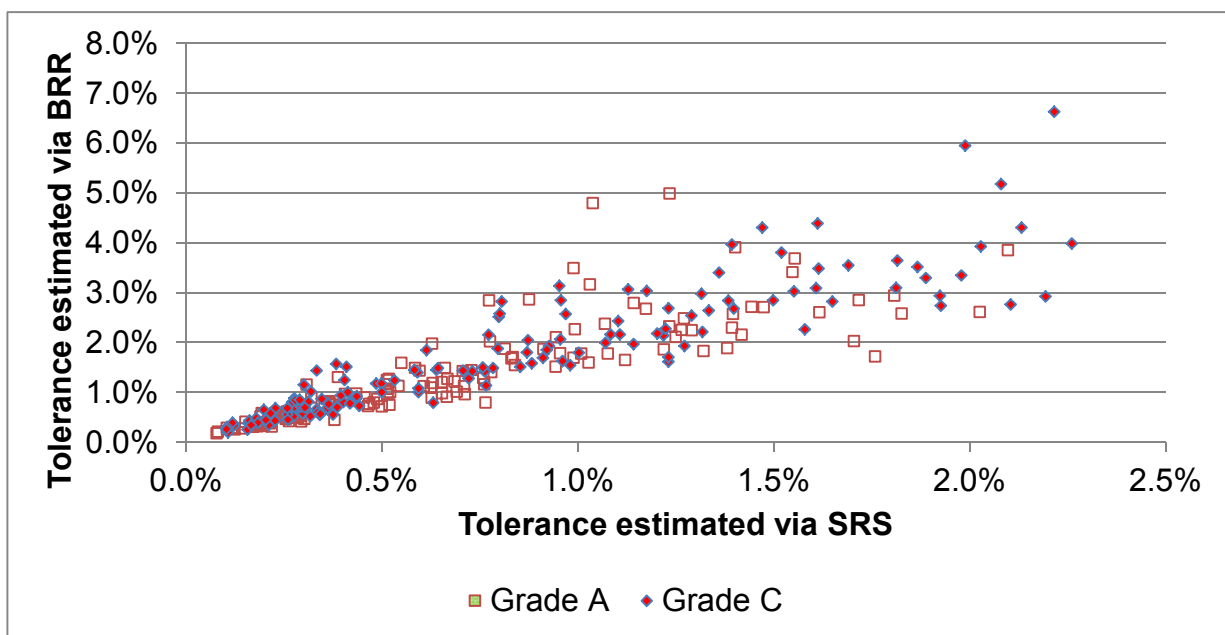
3.2 Comparison with tolerances calculated using simple random sampling (SRS) methods

An attempt to generate tolerances can also be undertaken using standard statistical sampling theory. If each candidate has a probability P of correctly achieving a given grade or above, then in a fully random sample of N candidates the standard error of the percentage of candidates that would actually achieve this grade or above within this sample is given by 100 times the square root of $P*(1-P)/N$. By substituting P with the proportion predicted to achieve a grade of interest or above and N by the number of candidates we can then use the normal approximation to estimate that an appropriate tolerance would be given by 1.15 times this amount.

There are a number of problems with this procedure. Firstly, candidates are not usually assigned to AOs on an individual basis but rather a whole centre will assign all their candidates to a single AO within any subject. Thus, the calculations in the previous paragraph ignore the important influence of centres on results. Secondly, every GCSE prediction is made for a fixed level of (KS2) prior attainment. The simple calculations ignore this as they assume that prior attainment may vary between different samples as well as outcomes. Finally, the simple calculations ignore the fact that there may be error in the prediction itself. Notwithstanding these criticisms, the aim of this section is to examine the relationship between SRS estimates of tolerance and the estimates generated via BRR in the previous section.

A comparison of the two sets of estimated tolerance is given in Figure 3.3. As can be seen there is a very strong relationship between tolerances estimated via the simple formulae provided by SRS and those provided by the more complex BRR procedure (correlation of 0.85). However, it can also be seen that estimates from BRR are considerably greater than the estimates from SRS. Specifically, Figure 3.3 shows that the estimated tolerances from BRR tend to be roughly double the estimates that would be derived from the simple formula. The reasons for this are likely to be due to the fact that the simple formula ignores the effects of individual centres and fails to take account of the fact that there may be error in the original prediction itself. However, more importantly, the analysis here shows that a relatively good approximation to BRR estimates of tolerances can be generated using some very simple formulae.

Figure 3.3: A comparison of tolerances estimated via SRS and BRR.



The analysis in this section implies that a more finely grained approach to tolerance could be adopted based on doubling the estimated tolerances from simple random sampling. Such an approach would be preferable to the recommendations provided in Table 3.2 because:

- These estimates take account not only of the number of candidates entering a GCSE but also of the proximity of the prediction outcomes to 50 per cent.
- These estimates avoid step changes in the recommended tolerances, thus avoiding the related issues explored earlier in Figure 3.2.

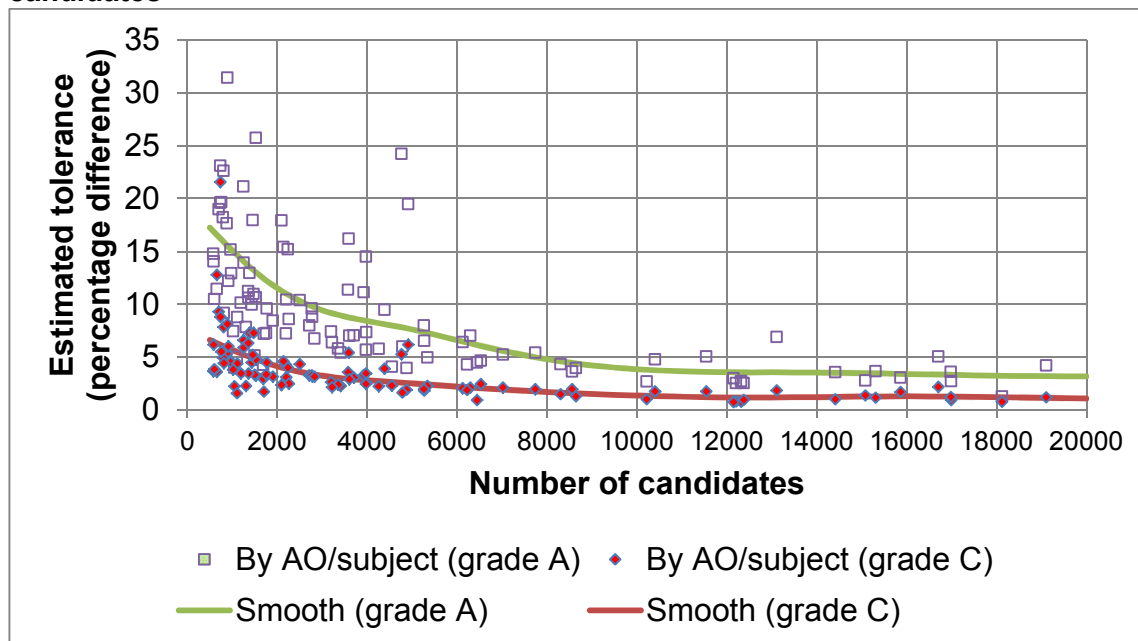
3.3 Quantifying tolerances as percentage rather than percentage point changes

At present tolerances are defined in terms of percentage point differences. That is, if we predict that 50 per cent of candidates will achieve C or above and then we see that in fact 55 per cent of candidates achieve C or above we say this is a difference of 5 percentage points. Another way to quantify differences would be in terms of percentage differences. That is, if we predict that 50 per cent of candidates will achieve C or above and then we see that in fact 55 per cent of candidates achieve C or above, we could note this as a difference of 10 per cent as the number of additional candidates who have achieved C or above is equal to a tenth of the number that were originally predicted to predict C or above. A potential advantage of considering tolerances in this way would be that it would ensure that we pay attention to differences from prediction where the overall number of candidates expected to achieve a particular grade is small. For example, in some subjects we might predict only very small numbers of candidates to achieve A or above and a difference of 1 percentage point may feel more important in this context.

However, in order for using tolerances in terms of percentages (rather than percentage points) to be a sensible approach it is desirable that it provides a relatively consistent approach to identifying tolerances across grades and across different subjects. This is explored further in this section. Using the same method described in Section 3.1 (BRR) we have calculated the estimated tolerances for each AO/subject combination in terms of percentage rather than percentage point differences. The results of this analysis are shown in Figure 3.4 which plots each of the estimated tolerances at grades A and C against the number of candidates taking an AO/subject combination. For the sake of consistency with Figure 3.1 the biggest outlier in the graph (WJEC Humanities with an estimated tolerance of almost 56 per cent) has been removed.

Figure 3.4 shows that estimated tolerances are highly inconsistent between grade A and grade C. For grade A the average tolerance is around 15 per cent for small subjects dropping to around 3 per cent for the largest subjects. In contrast, at grade C the average tolerance is around 5 per cent for the smallest subjects dropping to around 1 per cent for the largest. Furthermore, there is no obvious indication within either grade that quantifying tolerances in this way leads to a greater degree of consistency between AO/subject combinations. The inconsistencies between subjects and AOs for these estimates are caused by the fact that tolerances viewed in this way are highly dependent upon the number of candidates who are predicted to achieve the grades of interest. The smaller the predicted number, the larger the tolerance will be. This makes it difficult to provide simple rules for the most appropriate tolerance for any award dependent upon the candidates. More complex calculations to take account of this could be completed; however, these would be more complicated than the calculations recommended earlier in Section 3.3. For this reason we would not recommend that percentage differences were used as an alternative to percentage point differences to define recommended tolerances.

Figure 3.4: Relationship between estimated tolerances as percentages and number of candidates



3.4 Expected difference with screening predictions

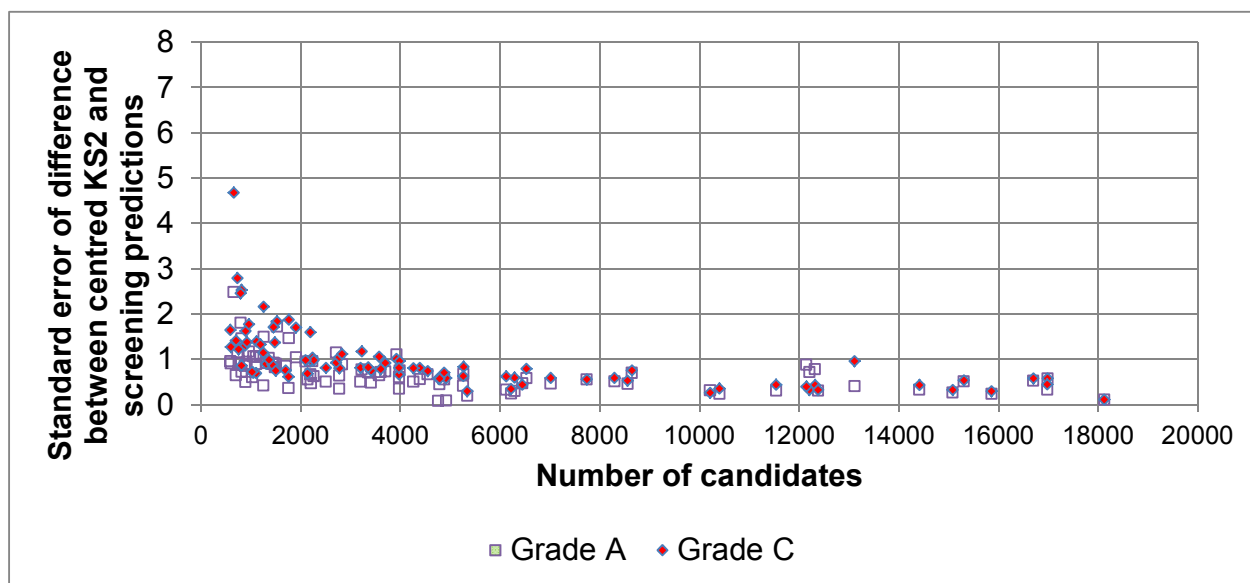
Although predictions based on KS2 may be useful to ensure inter-board comparability, a more powerful method to analyse differences between AOs is provided by the use of concurrent GCSE attainment (screening). The aim of analysis in this section is to explore the expected value of differences between predictions based on KS2 and those based on concurrent attainment.

Because we are interested in the value of screening for controlling inter-board differences, this section focusses on centred predictions. That is, within a given subject, the difference between the predicted national level of achievement and the predicted outcomes for each AO. All analysis was based on candidates with available KS2 data⁴². Analysis was again undertaken using BRR to calculate the variability in the difference between centred predictions based on KS2 and based on concurrent GCSE. The standard errors of these differences are presented at both grade A and grade C for each subject/AO combination in Figure 3.4. Because the focus is on inter-board differences, only subjects with at least two AOs with at least 500 candidates are included. As with Figure 3.1, AQA Environmental Science is not included in this plot and neither are AO/subject combinations with more than 20,000 candidates.

The results in Figure 3.4 show that the difference between KS2 and concurrent attainment shows little variation between different samples of centres. The standard error of the difference is less than 2 percentage points for nearly all subjects, and if the number of candidates is greater than 2000 it is nearly always less than 1 percentage point. This implies that where large differences are found between centred KS2 and screening predictions they are unlikely to be explained by random variation alone. This finding fits with analysis shown earlier (see Section 1) showing a relatively high correlation between centred predictions from KS2 and from screening. Having said this, the results also show that a small amount of difference between KS2 and screening predictions will be purely due to random fluctuations between samples.

⁴² Such candidates were overwhelmingly likely to also have available concurrent GCSE data. For this reason (and in order to allow the computationally intensive process of BRR to be used simultaneously for multiple purposes) analysis was not restricted to also only include those candidates with sufficient concurrent GCSE data. Thus, the two sets of centred predictions are based on ever so slightly different populations but this will not have a noticeable impact on the results presented in this section.

Figure 3.5: Standard errors of differences between centred KS2 and screening predictions



This analysis has shown that difference between centred predictions from KS2 and those from concurrent attainment concurrent attainment are stable across different samples of centres. However, they are not identical: concurrent attainment will provide slightly different predictions of where the outcomes for different AOs should sit against the national average than those provided by KS2. This issue will be explored further in the next section.

3.4 Summary

Analysis in this section has shown:

- Currently recommended tolerances underestimate the likely amount of variation between different samples of candidates with equivalent prior attainment. Particularly at grade C these tolerances should be adjusted upwards for future use.
- The current step changes in the recommended tolerances mean that many correctly awarded GCSEs could be judged as out of tolerance. A more finely grained system of tolerances may help to address this.
- A simple formula based on an adjusted version of the usual simple random sampling formula used to create confidence intervals could provide an improved mechanism to generate tolerances.
- There is no benefit to be gained from considering tolerances in terms of percentage differences from expectations rather than percentage point differences.
- Only a small amount of variation between prediction from KS2 and prediction from concurrent GCSE attainment is likely to be caused by random fluctuation. The relationship between screening and KS2 predictions will be explored further in the next section.

One issue we have not explored is how inter-board comparability would be strongly maintained in the context of increased tolerances. Specifically it is not clear how Ofqual could ensure that AOs apply consistent decision processes within this context and how any appearance of a 'race to the bottom' could be avoided within the tolerance levels recommended by this report. This issue will require ongoing discussions between Ofqual and the AOs.

4. Review of differences between screening outcomes and predictions

Earlier sections have begun to examine the differences between outcomes predicted using prior attainment at KS2 and those predicted using concurrent attainment. This section examines the relationship between the two further. Because the focus of this section is on the relative use of the two techniques to control inter-board differences, all predictions will be centred. That is, we will consider the predicted difference between each AO's outcomes and the national average. For this reason subjects with only one AO with more than 500 candidates will not be included in analysis. Only AOs with at least 500 candidates in any subject will be included in any figures.

4.1 Comparison of KS2 and screening predictions

As noted in Section 2.6 there is some evidence that KS2-based predictions underestimate the true extent of inter-board differences⁴³. This would mean that AOs with generally high attaining candidates (in terms of their other GCSEs) will end up with predicted outcomes that are too low whilst those with generally low attaining candidates will end up with predicted outcomes that are too high. This issue is explored further within this section alongside a more general consideration of the differences between KS2-based and screening predictions.

To begin with the predictions from KS2 and screening were recreated for all AO/subject combinations in 2013 using the data providing by awarding bodies (rather than the NPD). For this analysis, KS2-based predictions used the achievement of 16 year olds in 2011 and 2012 to predict achievement in 2013.

The results of these comparisons are shown in Figures 4.1 and 4.2. At both grade A and grade C a very strong relationship can be seen between the predicted performance (above the national average) based on KS2 and based on concurrent GCSE. This relationship is strongest when both calculations are undertaken based on the same set of candidates; those with matching KS2. In fact, in this analysis, we find a correlation of 0.90 between centred predictions from KS2 and those from concurrent attainment at grade C and a correlation of 0.89 at grade A. Similarly high correlations were found in the analysis of NPD data in Section 2.6. Furthermore, within the set of matched candidates centred predictions from KS2 tend to be very close to centred predictions from concurrent attainment. At grade C the average absolute difference in centred predictions is 0.8 percentage points with predictions from the two sources within 1 percentage point of each other 101 times (out of 137) and within 2 percentage points 124 times. At grade A the average absolute difference in centred predictions is even smaller at 0.6 percentage points and predictions from the two sources were within 1 percentage point of each other 106 times and within 2 percentage points in 132 times.

As might be expected if different sets of candidates are used to produce the two predictions then a greater degree of difference emerges. That is, if additional candidates with matching concurrent attainment but without matching KS2 data are used in screening then predicted achievement will change. Whilst there remains a relatively strong correlation between the two sets of predictions, (correlations of just above 0.7) it is clear that it is not necessary for them to match precisely.

⁴³ Further analysis has verified that the larger inter-board differences predicted by concurrent attainment are not caused by the fact that grades in individual subjects are included in the mean GCSE measure used to create them. This can partially be seen in the fact that a more complex use of GCSE attainment (see Section 4.2) leads to similar extent of predicted differences between AOs.

Figure 4.1: Differences between centred predictions from KS2 and concurrent attainment at grade C in 2013 (all candidates and matched candidates)

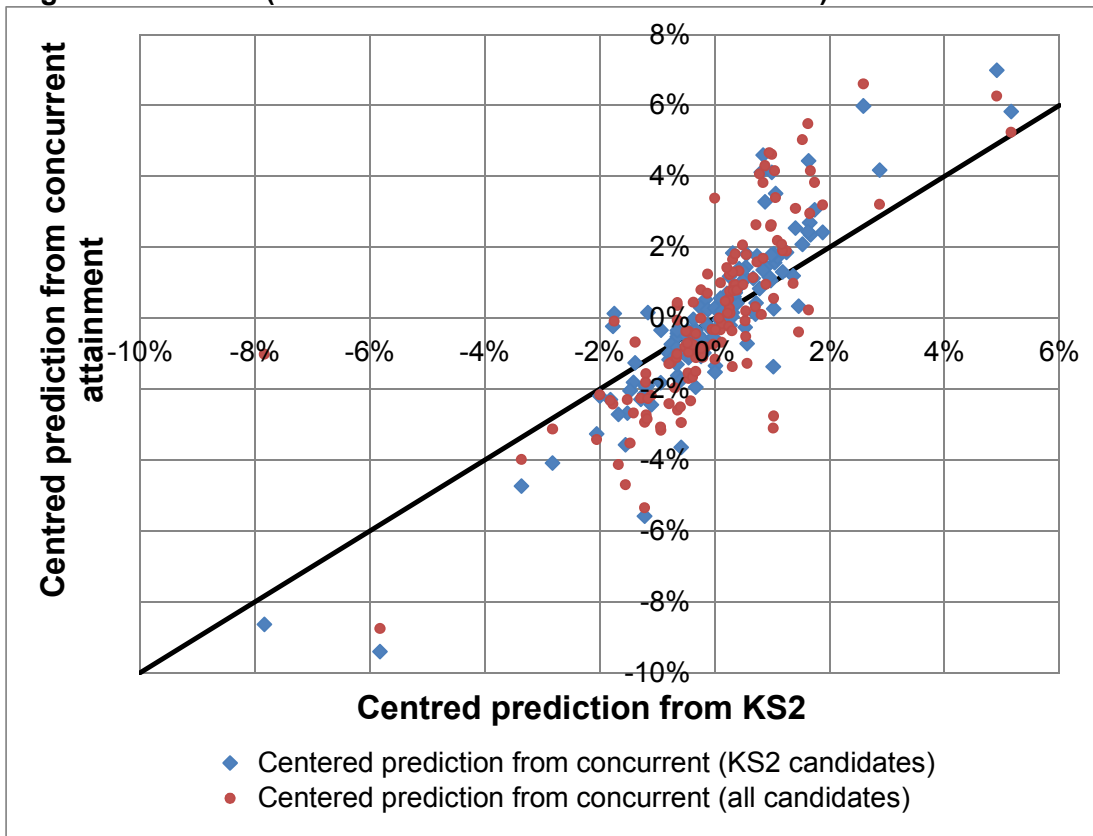
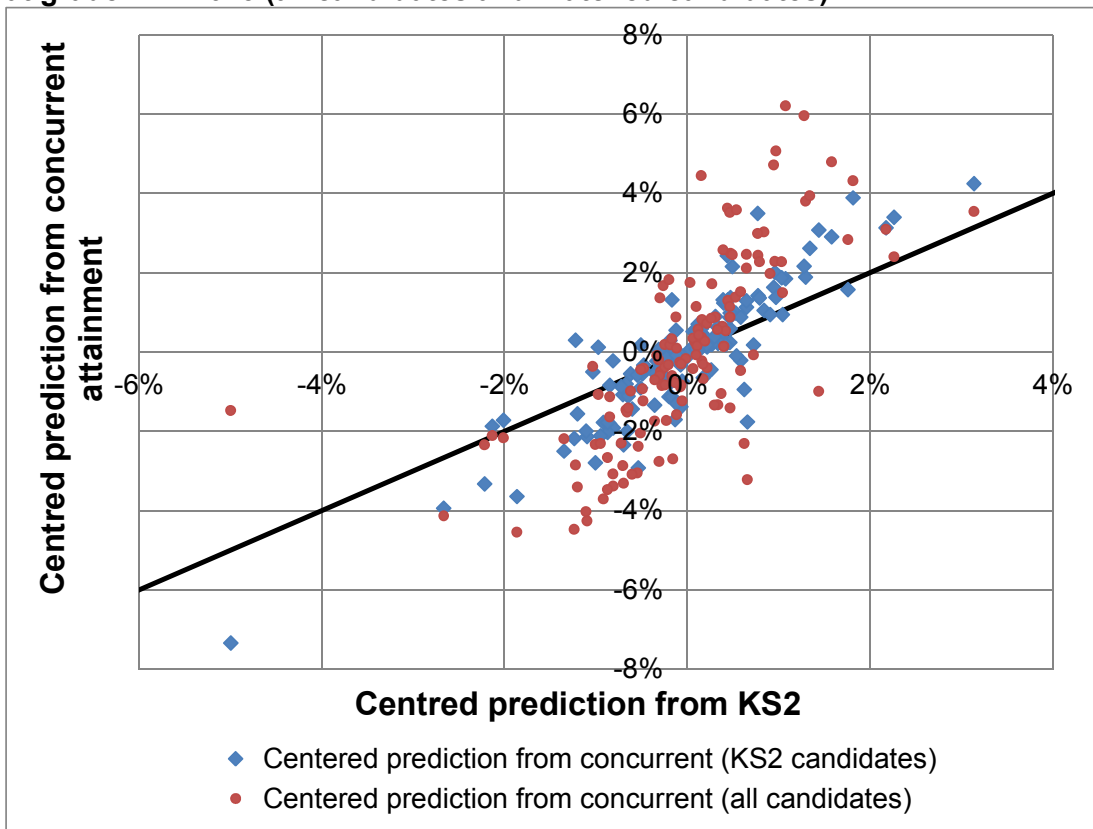


Figure 4.2: Differences between centred predictions from KS2 and concurrent attainment at grade A in 2013 (all candidates and matched candidates)



Although the results in Figures 4.1 and 4.2 are generally encouraging for the use of KS2 predictions, in that they are clearly very closely associated with predictions based on the (more powerful) measure of concurrent GCSE attainment, there are also causes for concern. In both Figure 4.1 and 4.2 a clear pattern for KS2 to marginally under predict differences between AOs emerges. A visual examination of these two charts suggests that, for the same set of pupils, if KS2 predicts that an AO's results will be 2 percentage points ahead of the national average, then concurrent attainment predicts that it will be 3 percentage points ahead. Similarly, it appears that if KS2 predicts that an AO's results will be 1 percentage point ahead of the national average, then concurrent attainment predicts that it will be roughly 1.5 percentage points ahead. Because a large number of the centred predictions are relatively close to zero, in many cases this apparent under-prediction makes little difference. However, in some cases the differences are more noticeable. Furthermore, given that a great many of the awards studied within these figures are encouraged to work within a tolerance of 1 percentage point (see Section 3), even these small differences may be of substantive importance.

4.2 Are screening predictions influenced by the combination of GCSE specifications candidates have taken at GCSE?

Before examining this effect further, it is first necessary to establish whether we can genuinely trust predictions from concurrent attainment as being more accurate than those from KS2. On the face of it we would assume that predictions based on concurrent attainment would provide a far more powerful tool to examine differences between AOs. This is chiefly because concurrent attainment is much more strongly correlated with attainment in any GCSE than achievement at KS2 (see Section 2.3). Furthermore, concurrent attainment allows us to predict the achievement of candidates based on their apparent ability at the time at which they are taking their GCSEs rather than 5 years previously (albeit in different subjects).

However, the use of concurrent attainment is not entirely unproblematic. Whereas at KS2 all pupils (within a cohort) will have taken exactly the same tests in exactly the same subjects, at GCSE pupils will have taken different combinations of GCSE subjects each set and awarded by different AOs. Even within the same AO there is often more than one available GCSE specification for the same subject. Thus at the very heart of using concurrent GCSE lies a potential problem; we need to assume comparability of GCSEs before we begin, and yet we cannot be sure about comparability of GCSEs until the process is completed. If the starting assumption is incorrect then the results may be flawed.

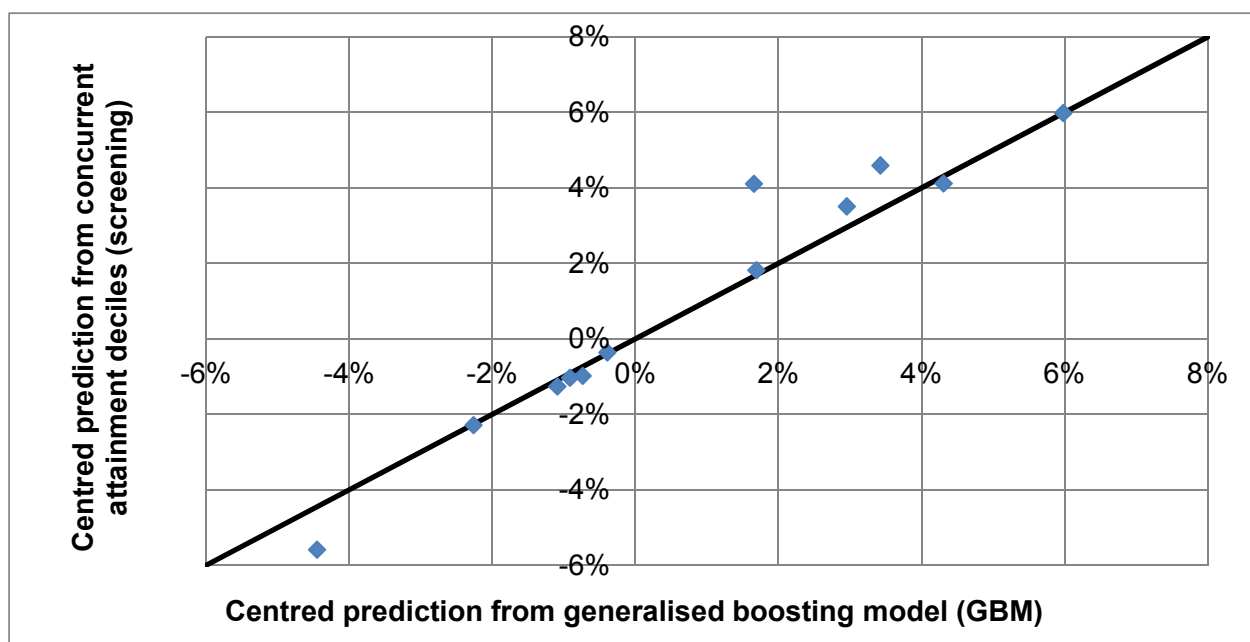
To further investigate this issue, predictions from the usual application of concurrent attainment were compared to a far more complex method based upon Generalized Boosted Models (GBMs, see Ridgeway 2012). These models work by combining many small predictive models to generate overall predictions. In our own scenario this is particularly valuable. Suppose we are interested in predicting the probability that a student will achieve a C or above in Music. We now build many small models examining how the chances of this event relate to achievement in different specific GCSE specifications. One model may examine the relationship with achieving grade A or above in Edexcel's Mathematics specification 1MA0, another may examine the relationship with achieving grade D or above in OCR's Geography specification J380. Crucially each of these models will include separate effects for if a candidate hasn't entered the particular specifications being used for prediction with each small model. By combining several thousand of these small models we can build up a very powerful overall predictive model that accounts, not only for candidates' achievement in other GCSE subjects, but also which specifications they studied and with which AOs. The specifications chosen to build each of the smaller predictive models and the overall number of small models that should be combined are both automatically optimised by the algorithms developed as part of GBMs. Crucially the predictions from these models are now based on achievement in precise GCSE specifications rather than a simple aggregation of grades across different subjects. Thus we can be confident that predictions are not compromised by differences in the comparability of different subjects and AOs at GCSE.

Fitting a GBM is a computationally intensive procedure. For this reason analysis was restricted to the 4 subjects showing the largest discrepancies for AOs with at least 1000 candidates, between KS2 and screening predictions; Citizenship Studies, D&T: Electronic Products, D&T: Food Technology, D&T: Product Design. Although there might only be a large discrepancy between KS2-based and screening predictions for one of the AOs offering this subject it was necessary to include data from all AOs in analysis in order to help build the predictive model. Because analysis was focussed on instances where there was a large discrepancy between KS2-based and screening predictions only candidates with matching KS2 data were included within analysis. To make the process computationally feasible the GBMs were constructed using GCSE grades in the 100 most popular GCSE specifications (across all subjects) in June 2013.

Predictions from GBMs are compared to predictions from the standard screening procedure in Figure 4.3. As can be seen there is a very strong association between the two sets of predictions. Only one point towards the centre of the graph stands out of showing any major discrepancy between the two predictions. This point relates to WJEC's D&T: Product Design GCSE. Only 780 matched candidates took this subject with WJEC and so it is likely that the difference between the methods is simply the result of random fluctuation.

Crucially we can see that there is no obvious tendency for predictions from screening to over-predict inter-board differences relative to GBMs. That is, the issue noted earlier with KS2 under-predicting the differences between AOs is not repeated with GBMs. On the contrary, both GBMs and the standard screening methodology predict similarly sized differences between AOs.

Figure 4.3: Comparison of centred predictions from screening and from GBMs



The results of this analysis show that differences between screening and KS2-based predictions are unlikely to be caused by any weakness in the assumptions of the screening methodology. Therefore, these differences are likely to indicate weaknesses in the current use of KS2 to predict GCSE results. Having accepted that such weaknesses exist, it is important to attempt to understand the cause of such weaknesses and to identify possible solutions.

4.3 Possible solutions to the issue of under-prediction of AO differences

4.3.1 Adjusting the KS2 method using ideas from equating

The issue of ensuring comparability between examinations falls into the general research topic of *equating*. Equating actually goes further than is required for GCSE standard maintaining. If we have two alternative versions of a test, then for every score on the first version, equating attempts to identify the exact score on the second version that is equivalent. Although equating and GCSE standard maintaining have slightly different aims⁴⁴, the techniques used are very similar. The equating technique most similar to that used at GCSE is *frequency estimation equipercentile equating*. This technique is broadly the same as that used to set GCSE grade boundaries. However, rather than using prior attainment (KS2) to link the two forms of a test, the link is established instead using the scores candidates have achieved on an anchor test.

Recent research has identified problems in the frequency estimation (FE) technique for equating. Specifically recent research (Wang and Brennan, 2009) has suggested that the basic assumption of the method may not hold in all circumstances. If we denote the scores candidates achieve on a particular test of interest as X, and the scores that candidates achieve on an anchor test as V, the basic assumption of the FE method is that the conditional distribution of X given a particular score achieved on an anchor test is invariant across different populations. This assumption is flawed because it fails to take account of measurement error in the anchor test itself. This means that in a low attaining population, any given anchor test score is likely to relate to a lower level of real ability than the same anchor test score in a high attaining population.

The fundamental assumption of the way in which KS2-based predictions of GCSE attainment are created is very similar to the assumption of the FE method; that is the probability of achieving a particular GCSE grade or above given a pupil's prior attainment is invariant between years and between AOs. However, if we imagine firstly that the AO a candidate is assigned to is more strongly related to their actual ability at the time of testing than to their KS2 results⁴⁵, and, secondly, that KS2 results are only an indicator of differences in ability between groups⁴⁶, rather than an exact measure, then in a low attaining population⁴⁷ any given level of KS2 attainment is likely to relate to a smaller chance of achieving higher GCSE grades than the same level of KS2 attainment in a high attaining population⁴⁸.

As well as identifying this potential problem in the use of FE equating, Wang and Brennan (ibid) also suggested a possible solution via a modified frequency estimation method. Their solution cannot be applied directly to the problem of KS2-based predictions of GCSE attainment as it is set in a different context. Specifically, whilst their method is concerned with adjusting for measurement error only, in our context we seek to address the low correlation between KS2 and GCSE achievement more generally. Whilst this low correlation may be partially attributable to measurement error, in the main it will be caused by the fact that different constructs are being measured. However, an argument analogous to the one used in the Wang and Brennan paper can be applied here, as described in Appendix 3. The resulting, modified method requires replacing the prior attainment scores of pupils within each AO/subject in the current year with the KS2 scores associated with the same level of concurrent attainment in the reference year. An

⁴⁴ Broadly speaking, equating is concerned with identifying comparable performance rather than comparable outcomes. Furthermore, equating requires that each test is measuring exactly the same construct, whereas, at GCSEs the constructs may change slightly over time as qualifications are reformed. In addition to this, each AO may assess a different syllabus within the same subject.

⁴⁵ This would make sense because the choice of an AO is likely to depend largely upon the school they attend. Schools will differ not only in terms of the prior attainment of their intake but also in their relative value added between KS2 and KS4. Thus, the choice of AO may relate more closely to the ability of candidates at the time of testing than to the KS2 attainment of candidates.

⁴⁶ Given the relatively low correlations between KS2 and GCSE attainment shown in Section 2 this assumption would appear reasonable. Note that the problem goes beyond the issue of measurement error dealt with in the research by Wang and Brennan (ibid). Not only do KS2 scores contain measurement error (as do any educational assessments) but also they are measuring a different construct to the one that is ultimately of interest, in different subjects and at a timepoint 5 years before GCSEs are taken.

⁴⁷ That is, an AO/year combination attracting the lowest ability candidates for a particular subject.

⁴⁸ That is, an AO/year combination attracting the highest ability candidates for a particular subject.

alternative possibility is to replace the prior attainment scores of pupils studying a particular subject within each *centre* in the current year with KS2 scores associated with the same level of concurrent attainment in the reference year.

4.3.2 Controlling for centre-level attainment in predictions

The methodology suggested above relies upon a particular set of, potentially controversial, assumptions about the underlying causes of the under-prediction problem; essentially stating that, at an individual level, not all KS2 results should be treated equally. Ideally, we would like to avoid such assumptions. With this in mind, a more straightforward reading of the under-prediction problem is that the current methods have failed to adequately capture the totality of the relationship between KS2 and GCSE grades. That is, when we examine results between AOs it is clear that there is still a residual relationship between the KS2 achievement of candidates within an AO and the mean GCSE achievement of candidates within an AO. This implies that the models described in Section 2 are inadequately capturing the full relationship. Furthermore, since under-prediction is an issue regardless of which measure of KS2 is used, this indicates that this problem cannot be resolved simply by more complex analysis of the relationship between KS2 and GCSE at the pupil level.

A possible solution to this issue is to base predictions not only on the KS2 achievement of each individual pupil, but also on the average level of achievement in their centre. The hypothesis here is that pupils with a given level of prior attainment will tend to achieve higher grades in a centre where the average KS2 level is high and lower grades in centres where the average KS2 level is low. Such “compositional effects” are researched extensively in the literature and are subject to some debate⁴⁹ as to their causes. However, in our own context, we are not interested in understanding the reasons for such effects or whether they are genuinely causal or not. Our only aim is to examine whether accounting for such effects in our models can aid the accuracy of predictions and help to address the under-prediction problem.

In order to control for centre-level prior attainment it is first necessary to calculate this value. For the purposes of this analysis, this is calculated as the mean normalised KS2 score of all of the pupils taking the given GCSE subject within a centre. If less than 5 pupils take the given subject within a centre, then the mean normalised KS2 score of all pupils taking the subject with the AO is used instead⁵⁰. Predictions are then made using logistic regression at each grade; controlling both for the normalised KS2 scores of individual candidates and for the mean KS2 score within their centre.

4.3.3 Using historical differences to adjust predictions

Using a similar logic to above, we might conclude that if a centre’s achievement has been under-predicted in one year then it is likely to be under-predicted in the following year. Therefore, if centres tend to remain with the same AO across years, any under-prediction (at AO level) in one year is likely to be carried forward to the next year to the same degree. Thus, an alternative to the modifications detailed above is simply to adjust each KS2-based prediction based on the historical level of difference between centred KS2 predictions and centred screening predictions. This adjustment has the potential advantage that it can potentially address numerous weaknesses in the prediction model at centre level in addition to the generic issue of under-prediction in which we are interested. For example, suppose an AO attracts centres with high value-added⁵¹, such that predictions from concurrent attainment tend to be higher than KS2-based predictions. Adjusting for historical differences between KS2 and screening predictions can take this fact into account without needing to identify a cause for the different levels of value-added.

⁴⁹ See, for example, Hutchison (2007) for more information.

⁵⁰ Replacing centre-level mean with the AO when there were less than 10 candidates for a subject within a centre was also trialled. This was found to make very little difference to predictions so is not discussed further within this report.

⁵¹ That is, the extent to which their pupils outperform expectations across all GCSE subjects.

In contrast to the other two solutions, this approach can only be applied if centred historical predictions are available. That is, if less than two AOs entered candidates in sufficient numbers in the previous year⁵² then no centred predictions will be available for the previous year and it will not be possible to apply any adjustments.

Note that adjustments based on historical performance require different information to the results typically provided to AOs at screening in two important ways:

- They are based only on candidates with matching KS2 data (as screening data is based on candidates' concurrent GCSE attainment only)
- They include a calculation of how far each AO's predicted outcomes differ from the national outcomes (rather than how AOs outcomes differ from each other).

Therefore, some modifications would need to be made to the current screening process in order for results to formally feed into standard maintaining in the following year. This would include the recalculation of screening statistics based only on pupils with matching KS2 data and the provision of national predictions for each subject both based on KS2 and concurrent GCSE.

4.3.4 Evaluation of the different solutions

Analysis was undertaken to evaluate how effectively each of the three approaches addressed the problem of under-prediction. For each AO/subject combination, in each of 2011, 2012 and 2013, centred predictions were generated using each of the above three methods with the Wang-Brennan adjustments applied both at individual centre level and overall AO level. Centred predictions based simply upon KS2 achievement without further adjustments were also generated. Finally, centred predictions from concurrent GCSE attainment were produced as the standard against which each of the other sets of predictions could be compared. Only subjects where all three of the above methods could be applied were included in analysis. Specifically, this meant that only subjects where historical data on the most appropriate adjustments was available were examined.

Predictions for 2013 were based upon data from 2012, predictions for 2012 were based upon 2011, and predictions for 2011 were based upon 2010. All KS2-based predictions were generated using logistic regression combined with normalised KS2 scores. Partly this is because earlier analysis in Section 2 has already identified some advantages with using this method. However, it is also advantageous in that both the Wang-Brennan modification and additionally controlling for centre-level attainment is most easily applied if predictions are based upon a continuous measure of KS2 attainment such as normalised scores⁵³.

The average absolute difference between centred predictions based upon each of the KS2-based methods and centred predictions based on concurrent attainment for each AO/subject combination was calculated at each of grade C and grade A in each year. The results are shown in Table 4.1.

⁵² Or if insufficient candidates as a whole took a GCSE in the subject in the year before that.

⁵³ Though it is by no means impossible for it to be applied to situations where KS2 is quantified in terms of distinct categories.

Table 4.1: Average absolute differences between centred predictions based upon KS2 and based upon concurrent attainment

Grade	Outcome Year	Method of adjustment for under-prediction					Number of AO/subject combinations analysed
		Unadjusted centred predictions	AO level Wang-Brennan	Centre level Wang-Brennan	Including centre-level prior attainment in models	Adjusting predictions based on historical data	
		Mean absolute difference between (centred) KS2 and screening predictions (percentage points)					
A	2013	0.58	0.54	0.55	0.56	0.42	136
	2012	0.51	0.50	0.51	0.50	0.45	126
	2011	0.52	0.52	0.53	0.52	0.45	123
C	2013	0.77	0.68	0.67	0.68	0.56	136
	2012	0.66	0.62	0.62	0.61	0.64	126
	2011	0.70	0.72	0.71	0.65	0.63	123

The results in Table 4.1 show that the Wang-Brennan adjustment generally improves the match between centred predictions from KS2 and centred predictions from concurrent GCSE. At best, the mean absolute difference between KS2 and screening predictions reduces from 0.77 percentage points down to 0.68 at grade C in 2013. Whilst, this gain is extremely small when viewed as a whole, Figures 4.4 and 4.5 show that for AO/subjects with either very high or very low centred predictions there are noticeable differences in predictions, with the modified KS2 method generally bringing predictions into line with those from concurrent GCSE. One exception to this rule is WJEC Mathematics at grade C (on the extreme left hand side of the graph). However these predictions are based upon a relatively small number of matched candidates (1,564) and as such may be largely caused by the effects of random fluctuations.

The results for the Wang-Brennan method are less encouraging in 2011 and 2012 with the method of adjustment leading to no improvement overall at grade A in 2011 and a slightly worse match with screening predictions at grade C. In spite of this, such an adjustment might still be worth considering if there were no better option. It is only when the predictions from KS2 are unbiased and known not to systematically under predict or over predict differences between boards that the tolerances derived in Section 3 can be confidently applied. Without confidence that predictions are unbiased, derived tolerances should be both larger and asymmetrically distributed around the predicted value meaning that they would be harder to apply in practice. For this reason, successfully addressing the under-prediction problem is important even if it doesn't lead to a greater match between KS2-based and screening predictions.

It can also be seen from Table 4.1 that applying the Wang-Brennan adjustment to individual centres leads to no improvement over applying the method to AOs as a whole. Indeed further exploration of the data found that both approaches yielded very similar predictions. For this reason this method is not explored further here.

Figure 4.4: Comparisons of centred predictions for 2013 at grade C with and without the Wang-Brennan modification

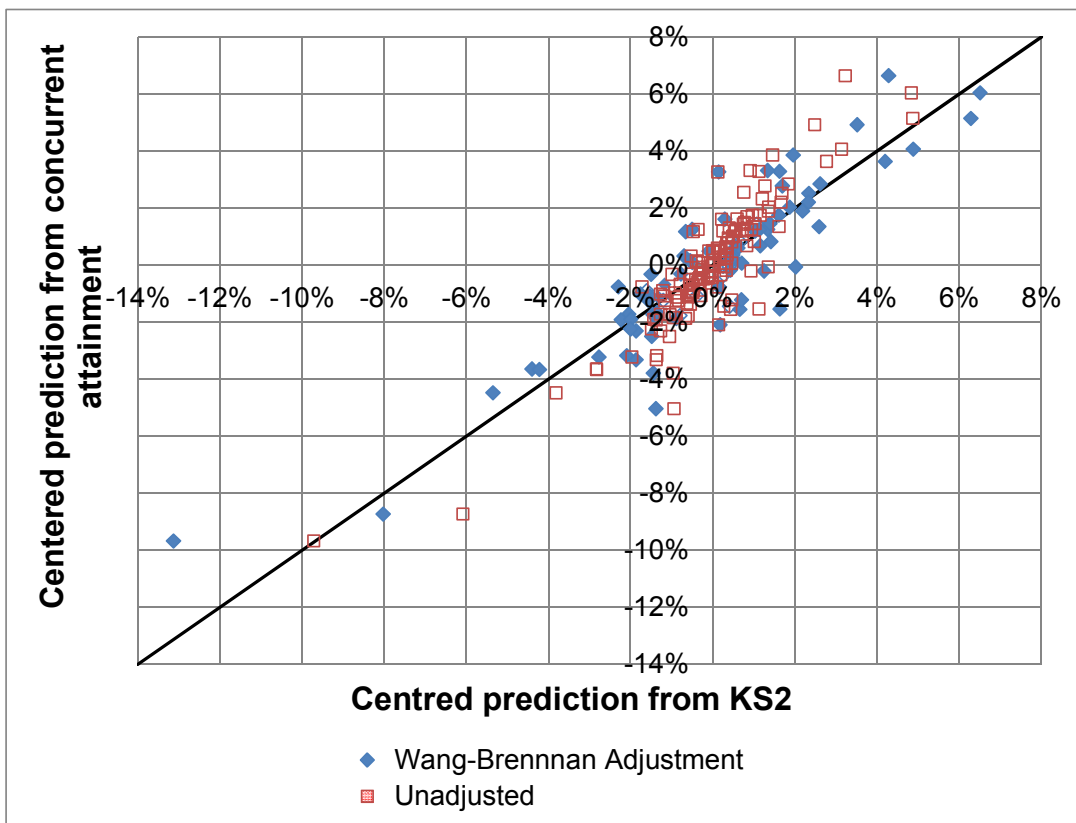
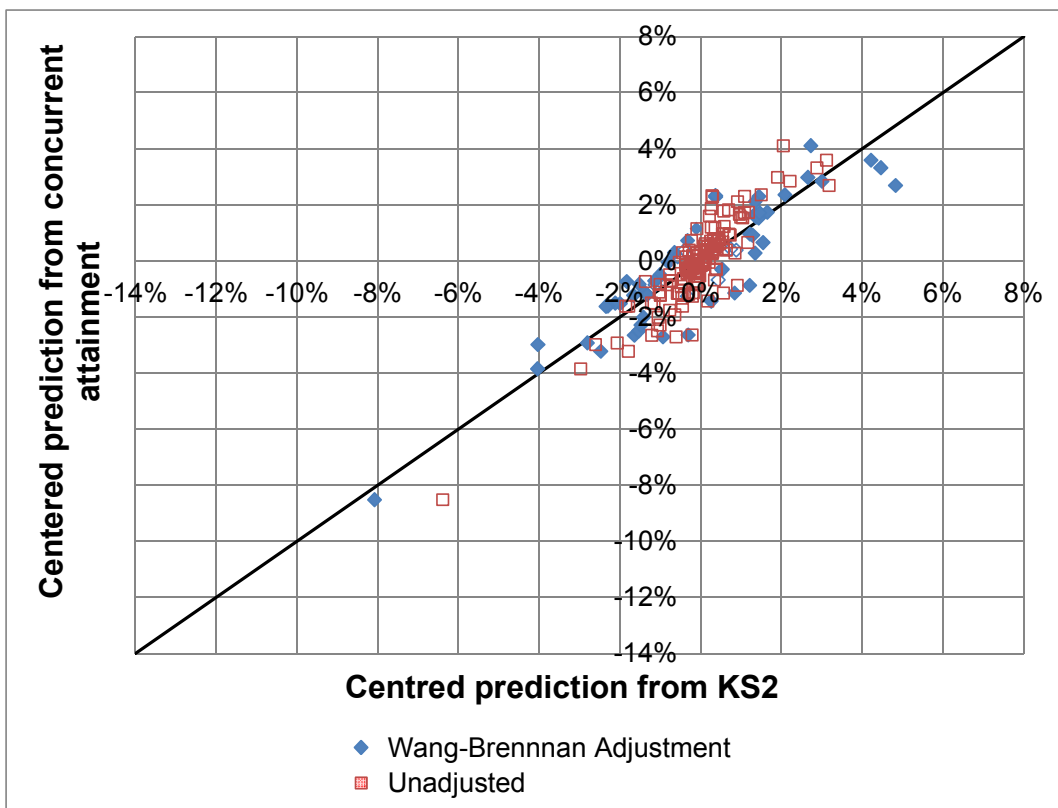


Figure 4.5: Comparisons of centred predictions for 2013 at grade A with and without the Wang-Brennan modification



Slightly more encouraging results are provided in Table 4.1 for additionally controlling for the average level of prior attainment within each centre. The table shows that predictions from this method are on average always at least as close to screening predictions as the unadjusted method, and, in fact, closer on average except for grade A in 2011. Further details, provided in Figures 4.6 and 4.7 also confirm that this method successfully addresses the issue of the under-prediction of differences between AOs. Furthermore, with the exception of grade A in 2013, the predictions from the method perform at least as well as the adjustments via the Wang-Brenan method and are closer for grade C in 2011 and 2012. This implies that controlling for centre-level attainment is a preferable approach to using the Wang-Brennan method.

Figure 4.6: Comparisons of centred predictions for 2013 at grade C with and without additionally controlling for centre-level prior attainment

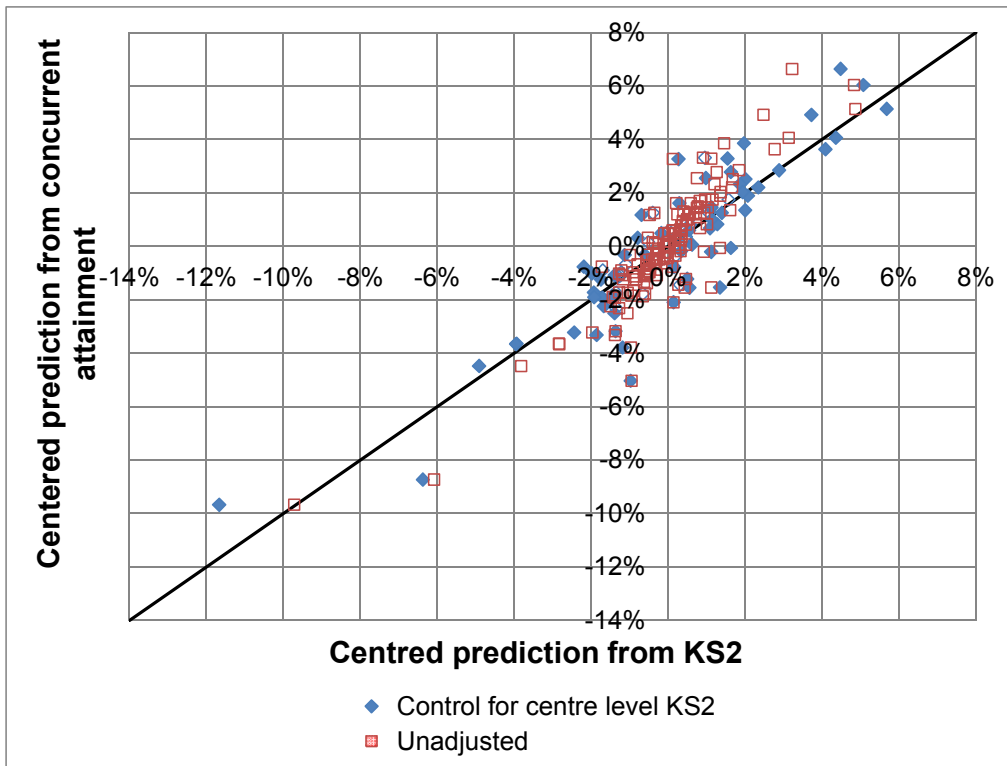
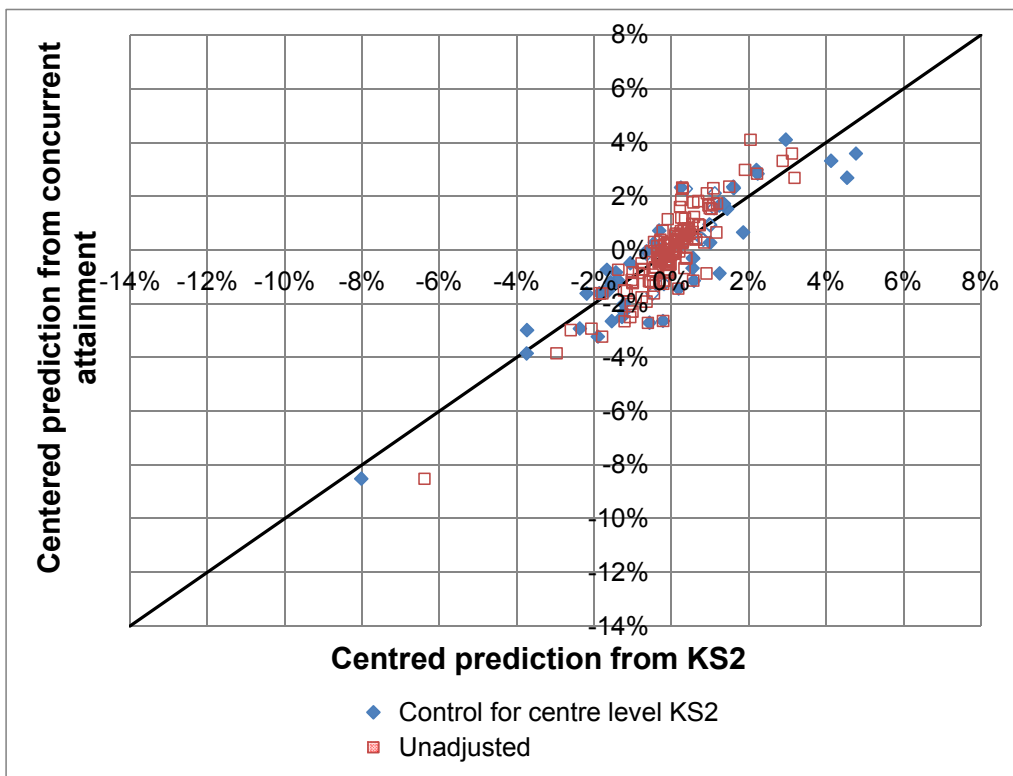


Figure 4.7: Comparisons of centred predictions for 2013 at grade A with and without additionally controlling for centre-level prior attainment



According to the results in Table 4.1, the method that leads to the closest match with screening predictions is to make adjustments based upon historical data. This result is further illustrated in Figures 4.8 and 4.9. The additional improvement compared to other methods is likely to relate to the historical differences addressing not only the issue of KS2 under-predicting differences between AOs, but also the fact that AOs are likely to retain many of the same centres between years. Centres with relatively high value added in one year (across all subjects) are likely to retain this high level of value added in the next year. Thus, adjusting for historical differences between KS2-based and screening predictions reduces not only the systematic bias in predictions but also extent of variation. This implies that, in circumstances where such adjustments are possible, this method provides the most appropriate means to address the issue.

Figure 4.8: Comparisons of centred predictions at grade C with and without adjustments using historical data

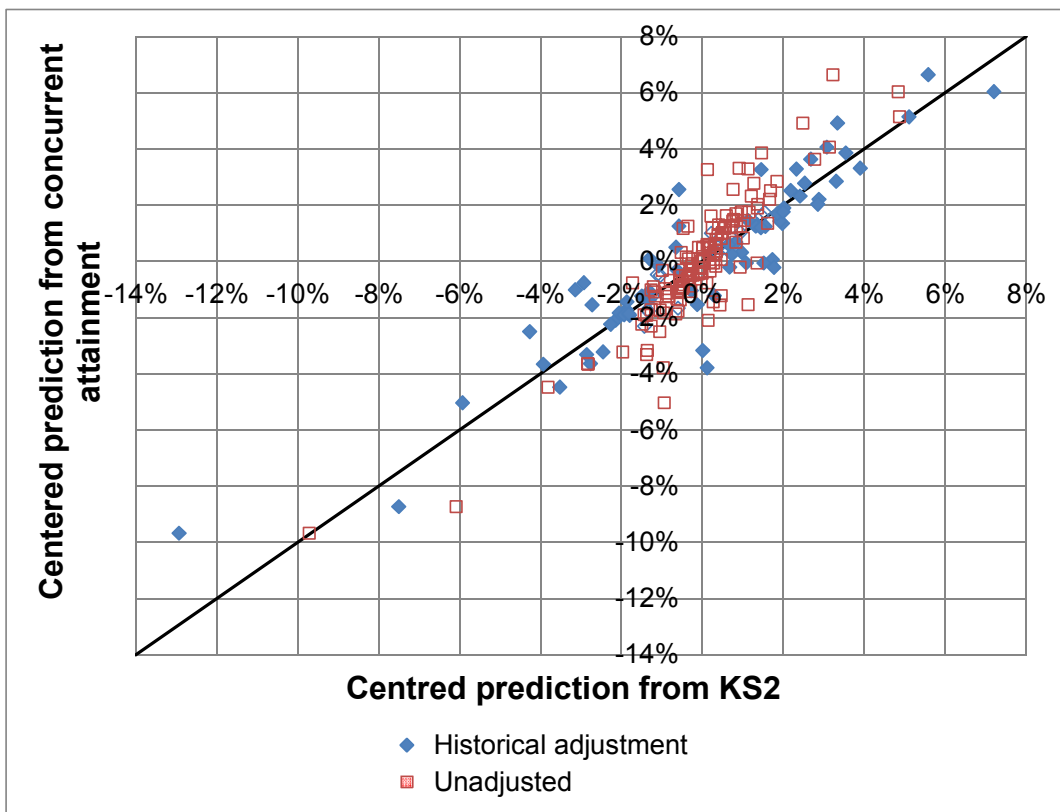
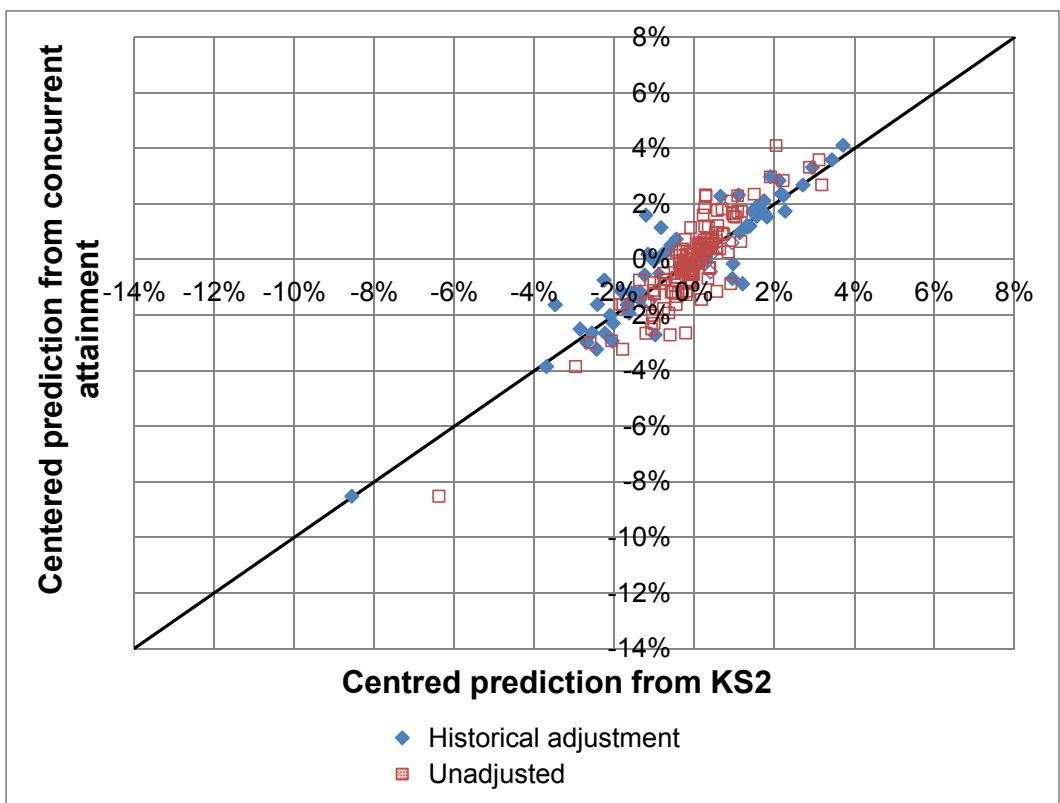


Figure 4.9: Comparisons of centred predictions at grade A with and without adjustments using historical data



4.4 Final thoughts on the under-prediction problem

Section 4.3 has identified three possible solutions to the issue of KS2 under predicting differences between AOs at GCSE: a modified method of calculation based on the research of Wang and Brennan (*ibid*), controlling for mean KS2 attainment within a centre or making adjustments based on historical differences in predictions. Further evaluation of the different methods shows that making adjustments based on historical differences between KS2 and screening predictions is the most effective of the different methods. Furthermore, it also has advantages in that:

- Assuming a relatively consistent allocation of centres to different AOs, it may help to control for the impact of individual centres.
- It does not require a fundamental change to the way in which calculations are done.
- It can be simply applied regardless of which measure of KS2 is used to produce - predictions. -

However, the results also show that additionally controlling for mean KS2 attainment within centres may also be effective. Although empirically it did not appear to be as effective as the method of using adjustments based on historical data it has the following additional advantages:

- It can be applied using information that is already available to AOs. It does not require any changes to the process of screening itself.
- As well as dealing to some extent with under-prediction of inter-board differences, it may address a potential similar problem with under-prediction of inter-year differences where the characteristics of pupils taking a particular subject changes dramatically. Whilst we have not been able to provide any empirical evidence on this issue⁵⁴, it would appear likely that such issues exist.
- It is robust in the event of large numbers of centres switching between AOs whereas the method based upon historical data relies on the fact that this has not happened.

Given the above advantages and disadvantages we would recommend that, in general, historical differences between screening and KS2-based predictions are taken into account.

However, in particular circumstances, such as where we expect larger than usual churn of centres between different AOs, controlling for mean KS2 attainment within a centre is likely to provide a more appropriate set of predictions. Indeed, in future it may be possible to combine both approaches, that is, augment the KS2-based predictions using centre-level information and also monitor the difference between these predictions and screening predictions over time. However, this possibility is beyond the scope of the current research.

Our analysis suggests that either of the above approaches provide more appropriate predictions than a modified Wang-Brennan method.

It should be noted that all of the above approaches have been based on trying to bring KS2-based predictions into line with predictions based on concurrent attainment. However, it could be argued that, for precisely the same reasons that KS2 under predicts differences between AOs, concurrent GCSE will also under predict inter-board differences⁵⁵. However, with no consistent measure of attainment that is a better predictor than mean GCSE, it is impossible to empirically prove whether this is the case or not. This may be an important area for further research in the future.

⁵⁴ There is no external method of determining the correct level for year-on-year differences. Whilst it is possible to produce predictions using mean GCSE based on relationship found within a reference year, this is not necessarily applying exactly the same standard as implied by the use of KS2 (as it is based on a different population) and so cannot be said to be definitively superior to the KS2-based method when looking at overall subject level changes between years.

⁵⁵ It is also possible that such a problem may affect predictions of A level outcomes based upon mean GCSE.

4.5 Summary

The analysis in this section has shown that:

- Predictions based upon KS2 are fairly consistent with predictions based on concurrent attainment. However, KS2 tends to very slightly under-predict inter-board differences.
- Differences do not appear to be due to any weakness in the way concurrent attainments aggregate achievement across GCSE subjects. More complex methods of using this data, which do not require this assumption, result in very similar predictions.
- The issue of KS2 under-predicting inter-board differences can be addressed either through additionally controlling for the average level of KS2 achievement within each centre or by adjusting predictions based on differences between KS2-based and screening predictions historically.

5. Appropriate tolerances for predictions based on concurrent GCSE performance

An alternative to using KS2 to produce predicted grade distributions for each GCSE subject for every AO is to produce predicted grade distributions using concurrent attainment. Whilst using such an approach to inform live awarding is not without its challenges, not least because concurrent attainment is not finally defined until awarding is complete, this is an option that is currently being explored by the AOs.

The aim of the analysis presented in this section is to apply the methods described in Section 3 to derive new tolerances for GCSE predictions that would be applicable in the situation where such predictions have been generated using concurrent achievement rather than KS2. This section should not be taken to imply a recommendation that such an approach *should* be taken. This section merely provides some technical details that may be useful for reference if such an approach is pursued further in future.

The method used in this section is identical to that described in Section 3; namely balanced repeated replication (BRR). Tolerances were estimated for each GCSE subject awarded by each AO in June 2013. Estimates were based on predictions generated using the performance of 16 year old candidates in both June 2011 and June 2012. As the use of concurrent data means there is no requirement for candidates to have matching KS2 data, all available candidates were included within analysis including those in independent and selective schools. However, only candidates with results recorded in at least 3 full GCSEs were included within analysis.

As described in Section 3, tolerances were estimated to represent 75 per cent confidence intervals. This means that if we had an independent means of knowing the “correct” grade boundary for each subject for each AO, and further that the model underlying predictions was true, then the correctly awarded outcomes for any subject within any AO would be within tolerance of predictions three quarters of the time.

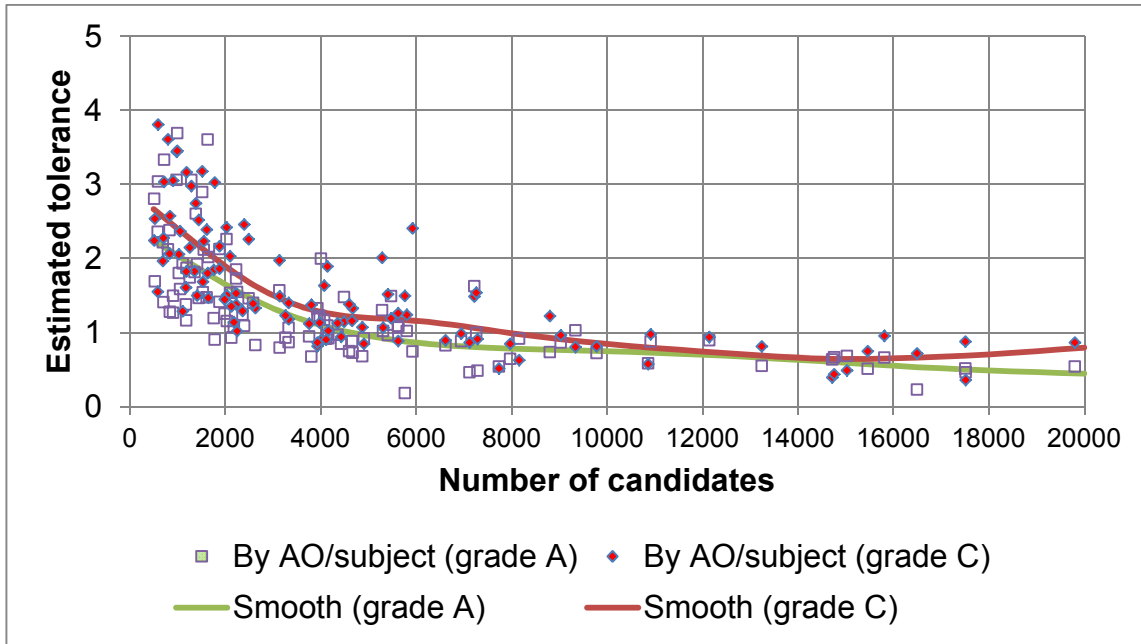
These tolerances are plotted against the number of matched⁵⁶ candidates taking each subject with the AO in Figure 5.1. Smooth lines showing how the average level of estimated tolerance at grades A and C changes dependent upon the number of candidates are included within each chart. As with the analysis in Section 3, one outlying subject (AQA Environmental Science) has been excluded.

In general Figure 5.1 shows a strong relationship between the number of candidates entering a qualification and the estimated tolerance. Furthermore, similar results are evident at both grade A and grade C, with generally smaller sample sizes required to yield the equivalent levels of reliability from KS2-based predictions. For example, as few as 500 candidates appear to be sufficient for the average level of tolerance to fall below 2.5 percentage points. This indicates that a tolerance level of 2 percentage points would be more appropriate than a tolerance level of 3 percentage points for this number of candidates. Similarly, the average level of tolerance falls below 1.5 percentage points at roughly 2500 candidates indicating a guideline tolerance of 1 will be more appropriate than 2 for this number of candidates.

Estimated tolerances at grade C tend to be slightly higher on average than those for grade A. However, the difference is slighter than was seen with KS2-based predictions and for larger sample sizes is lost amongst the general level of variation between subjects.

⁵⁶ In this case “matched” means those candidates with recorded results for at least 3 full GCSEs.

Figure 5.1: Relationship between estimated tolerances and number of candidates



Using the results above a set of possible set of appropriate tolerances for use with a concurrent GCSE prediction method is detailed in Table 5.1.

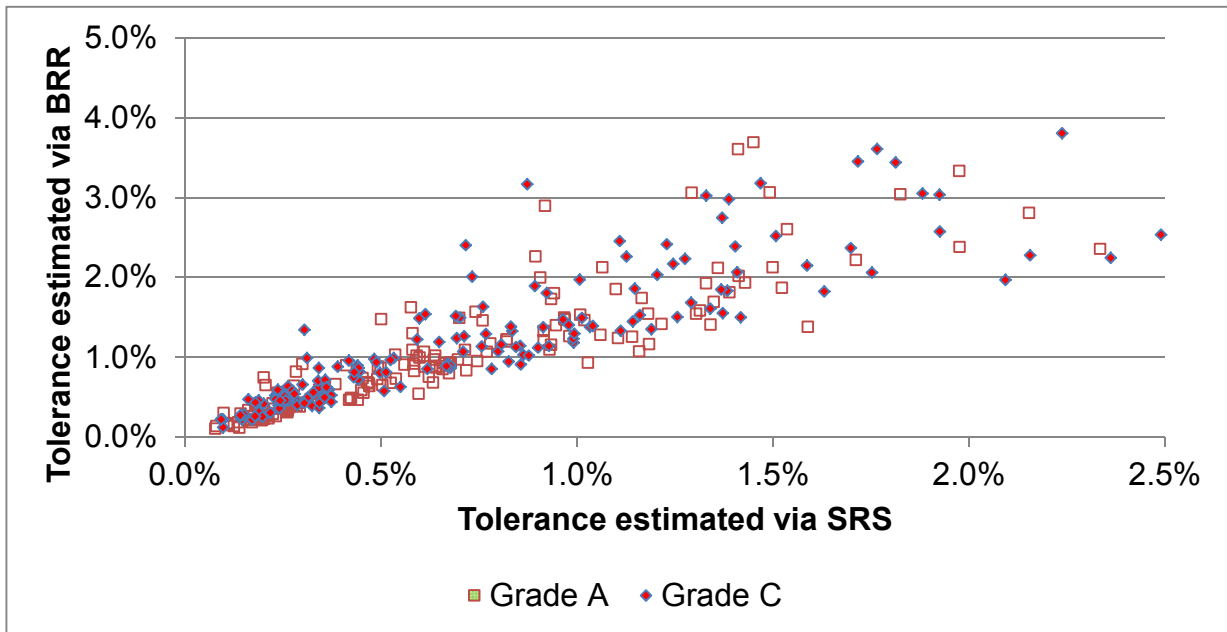
Table 5.1: Recommended tolerances for GCSE awarding based upon predictions using concurrent GCSE achievement

Recommended tolerance for predictions based on concurrent GCSE attainment	Sample size for tolerance to be applicable at each grade	
	Grade C	Grade A
2.5%	500-1500	N/A
2%	1501-2500	500-2000
1.5%	2501-4000	2001-3500
1%	4001+	3501+

5.1 Comparison with tolerances calculated using simple random sampling (SRS) methods

As described in Section 3.2 the tolerances derived via BRR can be compared to those derived using simple formulae based upon the assumptions of simple random sampling. A comparison of the two sets of estimated tolerances is given in Figure 5.2. As can be seen, there is a very strong relationship between tolerances estimated via the simple formulae provided by SRS and those provided by the more complex BRR procedure (correlation of 0.88). However, it can also be seen that estimates from BRR are consistently greater than the estimates from SRS. Specifically, Figure 5.2 shows that the estimated tolerances from BRR tend to be roughly 1.5 times as large as the estimates that would be derived from the simple formula. This is likely to be due to the fact that the simple formula ignores the effects of individual centres and fails to take account of the fact that there may be error in the original prediction itself. However, more importantly, the analysis here shows that a relatively good approximation to BRR estimates of tolerances can be generated using some very simple formulae.

Figure 5.2: A comparison of tolerances estimated via SRS and BRR for predictions based upon concurrent GCSE attainment



The analysis in this section implies that a more finely grained approach to tolerance could be adopted based on multiplying the estimated tolerances from simple random sampling by 1.5.

5.2 Summary

Analysis in this section has shown:

- Predictions based upon concurrent GCSE attainment are associated with smaller tolerances than those based upon KS2.
- A simple formula based on an adjusted version of the usual simple random sampling formula used to create confidence intervals could provide a simple mechanism to generate tolerances for such predictions.

6. Differences in the relationship between KS2 and GCSE achievement between years and AOs

6.1 Differences between years

An implicit assumption in the use of KS2 data to predict GCSE grade distributions is that the relationship between KS2 and GCSE remains stable over time. If this were not the case then a statistical model developed using data from one year may not provide the most appropriate predictions of outcomes in another year. The aim of this section is to explore the extent to which the relationship between KS2 and GCSE grades is stable for different GCSE subjects. Where differences are found it is also of interest to examine whether these differences can be explained by changes in the demographic characteristics of candidates taking a given subject. Finally we will examine the extent to which such changes make a practical difference to predictions.

To begin with, the relationship between GCSE grade and achievement at KS2 is examined in each year using multilevel modelling. For each GCSE subject the relationship between the total normalised KS2 score and GCSE grade is modelled including the following coefficients:

- An estimate of the average GCSE grade achieved by a candidate with a total KS2 normalised score of 100 (that is, with average KS2 achievement) in 2011.
- A single coefficient summarising the gradient of the association between normalised KS2 score and GCSE grade in 2011. Although more complex modelling may allow a more detailed assessment of the shape of the association between the two quantities, this approach provides a single estimate of the strength of the relationship between KS2 and GCSE. Furthermore, as has already been seen in Section 2.5, it is possible for simple models such as this to have superior predictive power to more complex approaches.
- Two coefficients estimating how the average GCSE grade achieved by a candidate with average KS2 achievement changes in 2012 and 2013 (the main effects).
- Two more coefficients estimating how the strength of the relationship between KS2 and GCSE grade differs in 2012 and 2013 from the relationship estimated in 2011 (the interaction effects).

In order to improve the speed of computation, GCSE grade was treated as a continuous variable and multilevel modelling was completed using the lme4 package in R⁵⁷. The multilevel aspect of the model was designed to account for the fact that centres may have an impact on the overall achievement of their students, and that the relationship between KS2 and GCSE grade may vary between centres. Furthermore, the multilevel model allowed for the fact that both of the aforementioned effects may vary between different years for the same centre.

For each subject, an additional model was fitted to the data including all of the coefficients above but also accounting for:

- The region in which each centre is located
- The level of deprivation in the locality where the candidate lives as measured by the Index of Deprivation Affecting Children (IDACI)
- The gender of the candidate
- An interaction between each of these characteristics and the relationship between KS2 achievement and GCSE grade.

Note that centre type was not accounted for in the models as, in common with the current practical application of prediction matrices, these models excluded candidates from independent or selective schools, and once these exclusions had been made the vast majority of candidates

⁵⁷ See <http://cran.r-project.org/web/packages/lme4/index.html>.

were within comprehensive schools meaning there was little point in accounting for additional centre types.

The crucial advantage of having used multilevel modelling for this analysis is that it allows us to assess the statistical significance of the differences between years both for the main and the interaction effects. Furthermore, the additional multilevel modelling allowed us to evaluate whether each of these effects remained statistically significant once the impact of changes in the demographic characteristics of candidates was taken into account.

The results of multilevel modelling are shown in Table 6.1. This table shows the numbers of statistically significant main and interaction effects identified within the multilevel models, both before and after, taking account of demographic information. A statistically significant main effect indicates that for a given subject in either 2012 or 2013, candidates with average KS2 achievement had significantly different (either higher or lower) achievement from similar candidates in 2011. A significant interaction effect indicates that in either 2012 or 2013 the strength of the relationship between KS2 and GCSE grade had significantly altered. For each of 53 subjects included in analysis, comparisons with 2011 were made both for 2012 and 2013. This means that a total of 106 effects were evaluated.

Table 6.1: Numbers of statistically significant main and interaction effects (at the 1% level) in multilevel models examining changes in the relationship between KS2 and GCSE grade over time

	After taking account of demographic factors					
	Main Effects			Interactions		
Before taking account of demographic factors	No	Yes	Total	No	Yes	Total
No	79	6	85	63	3	66
Yes	6	15	21	8	32	40
Total	85	21	106	71	35	106

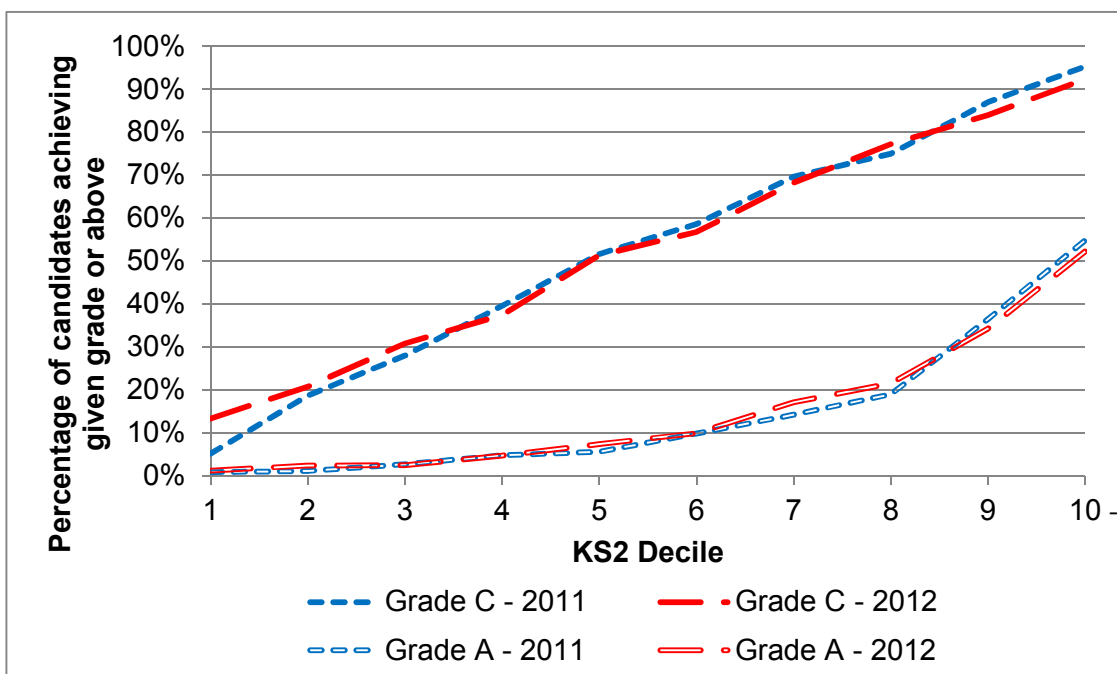
Table 6.1 shows that significant main effects were only found for a minority (21 out of the 106) of the cases studied. Furthermore, the number of significant main effects is not reduced by further accounting for other demographic factors. This implies a significant difference between the GCSE attainment of pupils with average levels of KS2 in different years. The fact that such effects are found, even where the use of KS2 data is specifically designed to remove them, could be caused by two things. Firstly, because grade boundaries are required to coincide with an exact, whole number of marks, it may not always be possible for AOs to place grade boundaries in such a way that awards will be precisely in line with predictions. Secondly, it should be remembered that AOs are only encouraged to award grades in line with predictions if “there is not sufficient evidence of real improvements in performance” (Ofqual, 2012). If such other evidence is available, AOs can award either fewer or greater numbers of higher grades as appropriate. Given these two points, and also that mechanisms to examine such differences already exist, the main effects are not of particular interest in this section.

Table 6.1 shows that a slightly larger number of statistically significant interactions (40 out of 106) were identified. This is to be expected since current methods of awarding specifically address possible differences in the relative difficulty of GCSEs in different years (the main effects), but do not (and possibly cannot) seek to maintain the strength of the relationship between KS2 and GCSE. After taking account of other demographic factors, the significance of the majority of these effects remained unchanged, although there was a slight reduction in the number of statistically significant interactions effects (from 40 to 35 out of 106). However, this does not necessarily imply that any change in the strength of the relationship between KS2 and GCSE grade was *caused* by demographic factors in these cases. The results may simply reflect

the increased difficulty of identifying changes in the relationship whilst accounting for a number of demographic characteristics⁵⁸.

Having concluded that the strength of the relationship between KS2 and GCSE grades has changed over time in at least some subjects, it is now of interest to examine the likely impact of these changes on predictions for individual AOs. This is illustrated using data from GCSE Psychology in 2012. This subject is chosen because, of all of the statistically significant interactions identified for 2012, this was the largest. That is, this is the subject that displayed the largest statistically significant change in the relationship between KS2 and GCSE between 2011 and 2012. The difference in the relationship between KS2 and GCSE between 2011 and 2012 is shown in Figure 6.1. For the purposes of this chart, in order to allow easy interpretation, KS2 achievement is categorised in terms of deciles and GCSE attainment is examined in the terms of the percentage of candidates achieving C or above and A or above. The same information provided in Figure 6.1 is also provided in Table 6.2.

Figure 6.1: Changes in the relationship between KS2 and GCSE Psychology between 2011 and 2012



As can be seen from Figure 6.1 and Table 6.2 there were some small differences between the achievement of pupils at each level of prior attainment in different years. The statistically significant interaction effect is manifested in that, at lower levels of prior attainment, candidates in 2012 tended to outperform candidates in 2011, whereas at higher levels of prior attainment the reverse was true. This implies that the relationship between KS2 and GCSE achievement in Psychology was slightly weaker in 2012 than in 2011.

⁵⁸ That is, the statistical *power* of the analysis is likely to be reduced once we try to account for other factors.

Table 6.2: Changes in the relationship between KS2 and GCSE Psychology between 2011 and 2012

KS2 Decile	Number of matched candidates		Percentage achieving C or above		Percentage achieving A or above	
	2011	2012	2011	2012	2011	2012
1	117	165	5.1%	13.3%	0.9%	1.2%
2	359	424	18.7%	20.8%	1.1%	2.4%
3	529	649	28.0%	30.8%	2.6%	2.5%
4	670	791	39.6%	37.4%	4.8%	4.7%
5	856	991	51.6%	51.4%	5.6%	7.4%
6	925	1138	58.6%	56.8%	9.8%	9.9%
7	1076	1164	69.6%	68.2%	14.2%	17.1%
8	1096	1269	75.0%	77.2%	19.0%	21.5%
9	1057	1200	86.9%	84.0%	36.4%	34.3%
10	874	1060	95.3%	92.4%	54.8%	52.3%

A possible effect of a decrease in the strength of the relationship between KS2 and GCSE, such as that shown above, is that predicted differences between AOs may decrease in the following year. This is investigated further in Table 6.3. This table shows how the predicted percentage to achieve at least grades C and A in Psychology in 2013 in each AO differs from the national prediction. These predictions are based upon historical data either from summer 2011 or 2012. As can be seen from the table, the choice of reference year makes very little difference to these centred predictions. For example, at grade C achievement within AQA (the market leader) is predicted to be 1.15 percentage points below the national average using data from 2011 and 1.06 percentage points below the national average using data from 2012. In other words, even in the subject with the largest identified statistically significant interaction coefficient, the predicted differences between boards are barely affected. The largest differences occur at grade C for WJEC. However, only 423 matched candidates are available in this case and so these differences would be dwarfed by the size of the uncertainty associated with these predictions.

Table 6.3: Centred predictions for each AO for Psychology GCSE in 2013

AO	Centred predicted percentage to achieve C or above based on...		Centred predicted percentage to achieve A or above based on...		Number of matched candidates (2013)
	2011	2012	2011	2012	
Edexcel	0.51%	0.46%	-0.22%	-0.18%	1840
WJEC	6.44%	6.02%	3.52%	3.38%	423
OCR	0.30%	0.27%	0.18%	0.17%	3505
AQA	-1.15%	-1.06%	-0.42%	-0.42%	4102

Having studied the subject with the largest interaction coefficient and found that the choice of reference year makes little difference to predicted difference between boards, we can conclude that changes in the strength of the relationship between KS2 and GCSE do not have important practical implications.

This does not imply that the choice of reference year is itself completely unimportant. As was seen in Table 6.1, a number of significant main effects have been identified. These imply that the use of different reference years is likely to make a visible difference to national results in a given subject. However, a number of mechanisms are already in place to control changes in the national standard over time (of which, KS2-based prediction is already one). This means that the choice of reference year remains an important judgemental decision, but not one that can be chosen purely on statistical grounds. Over time, ideally, the same reference year will be retained⁵⁹ and used to generate predictions for several subsequent years. This will in itself ensure that different main effects cannot influence predictions. The analysis in this section has shown that, at least in the short term, changes in the strength of the relationship between KS2 and GCSE are unlikely to have major practical implications. Having said this, this issue may be worth revisiting to explore differences over a longer time period.

6.2 Differences between AOs

It is also of interest to explore the extent to which the relationship between KS2 and GCSE differs between different AOs in the same year. In order to investigate this issue, the relationship was examined using multilevel modelling in a similar way to that described in the previous section. However, within each subject with at least two AOs⁶⁰, the analysis now focussed on whether there were statistically significant differences between the market leader⁶¹ and other AOs in their KS2-GCSE relationship. As in the previous analysis, both main effects⁶² and interactions⁶³ were considered and the analysis was run both before and after taking account of the influence of other demographic variables. Separate analyses were run for each subject in each of the years 2011, 2012 and 2013.

The results of multilevel modelling are shown in Table 6.4. This table shows the numbers of statistically significant main and interaction effects identified within the multilevel models both before and after taking account of demographic information. A statistically significant main effect indicates that for a given subject for a particular AO, candidates with average KS2 achievement had significantly different achievement (either higher or lower) than similar candidates studying the same subject with the market leader. A significant interaction effect indicates that for a particular AO the strength of the relationship between KS2 and GCSE grade was significantly different than was seen with the market leader. Note that, for each of 45 subjects included in analysis, comparisons with the market leader were made with every other AO with at least 500 candidates. This means that a total of 283 effects were evaluated.

Table 6.4 shows that significant main effects were found for just under a third (81 out of 283) of the cases studied. Furthermore, the number of significant main effects slightly increases (to 85) after further accounting for other demographic factors. The statistically significant main effects imply a significant difference between the GCSE attainment of pupils with average levels of KS2 studying with different AOs. As discussed previously, these effects could be because of the requirement for grade boundaries to coincide with a whole number of marks or because AOs are only encouraged to award grades in line with predictions if “there is not sufficient evidence of real improvements in performance” (Ofqual, 2012). Given these two points, and also that mechanisms to examine such differences already exist, the main effects are not of particular interest in this section.

⁵⁹ This had not yet been achieved for predictions in any of 2011, 2012 or 2013.

⁶⁰ With at least 500 candidates each in any year .

⁶¹ That is, the AO with the greatest number of matched candidates within a particular year. This is used as the comparator to other AOs within analysis because, being the AO with the largest amount of data, it provides the most reliable comparison.

⁶² That is, whether there are differences in the achievement of a pupil with an average level of prior attainment between the market leader and another AO.

⁶³ That is, whether there are differences in the strength of the relationship between KS2 and GCSE grade between the market leader and another AO.

Table 6.4: Numbers of statistically significant main and interactions effects (at the 1% level) in multilevel models examining differences in the relationship between KS2 and GCSE grade between the market leader and other AOs

	After taking account of demographic factors					
	Main Effects			Interactions		
Before taking account of demographic factors	No	Yes	Total	No	Yes	Total
No	187	15	202	192	12	204
Yes	11	70	81	13	66	79
Total	198	85	283	205	78	283

Table 6.4 shows that roughly a third of the differences studied (79 out of 283) yielded a statistically significant interaction effect. Taking account of other demographic factors has little impact on the number of significant interaction effects that were identified. As with the analysis of changes over time, significant interaction effects are to be expected since current methods of awarding specifically address possible differences in the relative difficulty of GCSEs in different AOs (the main effects), but do not seek to restrict the strength of the relationship between KS2 and GCSE.

Having seen that the strength of the relationship between KS2 and GCSE grades differs between AOs in at least some subjects in some years, it is now of interest to examine the likely impact of these changes on predictions for individual AOs. This is illustrated using data from GCSE Biology in 2012. This subject is chosen because, of all of the statistically significant interactions identified for 2012, this interaction effect examining the difference between Edexcel and AQA was one of the largest⁶⁴. That is, this subject displayed one of the largest statistically significant differences in the relationship between KS2 and GCSE between the market leader (AQA) and another AO. The difference in the relationship between KS2 and GCSE between AQA and Edexcel in 2012 is shown in Figure 6.2. The same information provided in Figure 6.2 is also provided in Table 6.5.

As can be seen from Figure 6.2 and Table 6.5, at lower levels of prior attainment, candidates in Edexcel tended to outperform candidates in AQA. However, at higher levels of prior attainment AQA's candidates tended to outperform Edexcel's. This confirms that the relationship between KS2 and GCSE achievement in Biology in 2012 was slightly weaker in Edexcel than in AQA, that is, the KS2-GCSE relationship is flatter for Edexcel than for AQA. This in itself suggests that, although the use of statistical predictions can enforce "comparable outcomes" at an aggregate level, it cannot necessarily ensure that two specifications are equally difficult for all possible different student types.

⁶⁴ Specifically it was the second largest. A slightly larger difference was found between OCR and AQA for Environmental Science but the relatively low numbers of matched candidates involved (1038 and 698 in OCR and AQA respectively) meant that this was not ideal for illustrative purposes.

Figure 6.2: Differences in the relationship between KS2 and GCSE Biology between Edexcel and AQA in 2012

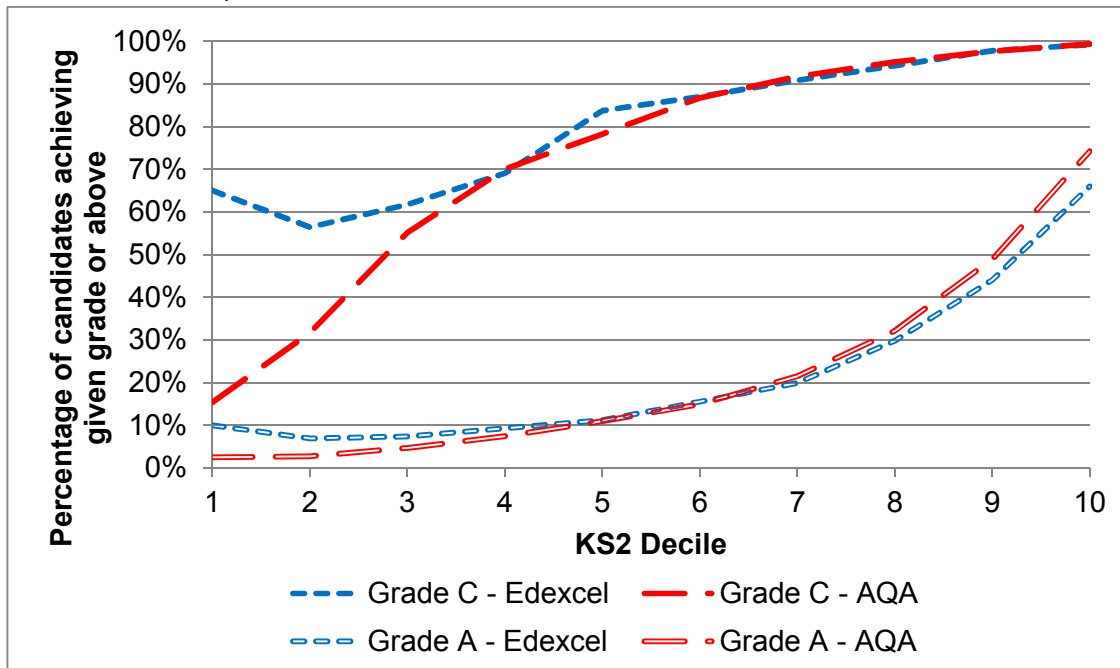


Table 6.5: Differences in the relationship between KS2 and GCSE Biology between Edexcel and AQA in 2012

KS2 Decile	Number of matched candidates		Percentage achieving C or above		Percentage achieving A or above	
	Edexcel	AQA	Edexcel	AQA	Edexcel	AQA
1	20	202	65.0%	15.3%	10.0%	2.5%
2	101	477	56.4%	31.7%	6.9%	2.7%
3	204	883	61.8%	55.2%	7.4%	4.6%
4	407	1718	69.0%	70.0%	9.3%	7.5%
5	708	2851	83.8%	78.3%	11.2%	11.0%
6	1083	4610	87.0%	86.7%	15.5%	14.9%
7	1568	6833	90.8%	91.7%	20.0%	21.5%
8	2152	9468	94.2%	95.2%	29.8%	32.1%
9	2644	11959	97.8%	97.6%	44.1%	48.9%
10	2930	13736	99.2%	99.4%	66.0%	74.2%

Of course, in subjects provided by multiple AOs, predictions would never be based upon historical results from a single AO. However, a theoretical possibility would be that, over time, the strength of the relationship between KS2 and GCSE became more like the relationship in one of the AOs. If this happened then, theoretically, predicted differences between AOs should be increased or decreased accordingly. The possible extent of this effect is investigated further in Table 6.6. This table shows how the predicted percentage to achieve at least grades C and A in Biology in 2013 in each AO differs from the national prediction. These predictions are based upon historical data from summer 2012 in either Edexcel or AQA. As can be seen from the table, in fact, the choice of which AO is used makes very little difference to these centred predictions. For example, at grade C achievement in 2013 within AQA (the market leader) is predicted to be 0.17 percentage points above the national average using data from Edexcel and 0.21 percentage points above the national average using data from AQA. In other words, even one of

the largest identified statistically significant interaction coefficients relates to hardly any change in the predicted differences between boards.

Table 6.6: Centred predictions for each AO for Biology GCSE in 2013

AO	Centred predicted percentage to achieve C or above based on...		Centred predicted percentage to achieve A or above based on...		Number of matched candidates (2013)
	Edexcel 2012	AQA 2012	Edexcel 2012	AQA 2012	
Edexcel	0.15%	0.18%	0.38%	0.46%	12436
WJEC	-0.52%	-0.58%	-3.69%	-4.20%	158
OCR	-0.32%	-0.39%	-0.89%	-1.06%	40011
AQA	0.17%	0.21%	0.50%	0.59%	63612

6.3 Summary

Analysis in this section has shown:

- In a minority of cases there are statistically significant differences in the relationship between KS2 and GCSE in different years and within different AOs.
- In general these differences do not appear to be caused by differences in the demographic characteristics of candidates. -
- Within the data analysed, differences in the strength of the relationship between KS2 and GCSE between years and between AOs are not large enough to have any practical impact on predicted differences between AOs. It may be worth continuing to monitor such differences over time to verify that this continues to be the case.

7 Further investigation of centre effects

This section further explores two effects related to the centres which candidates attend: firstly whether the current exclusion of candidates at selective and independent schools from predictions is justified and can be improved upon, and secondly how accounting for mean centre-level KS2 attainment, as proposed in Section 4.3 as a solution to the under-prediction of differences between AOs, affects the fit of the prediction model.

7.1 Using centre type in predictions

The current methods (and the comparisons drawn in Section 2 of this report) exclude candidates at selective and independent schools due to evidence of a different value-added relationship at these centres (Eason 2010). This difference would be of no importance if the distribution of centre type remained the same from year to year, as Eason points out, but at an individual subject and AO level, such an assumption is not realistic. An alternative way of accounting for this would be to generate predictions (for all candidates) that explicitly depend on centre type, thereby accommodating the variation and using it to make overall predictions more accurate. This section investigates the effect of doing so for total normalised marks (K2NrTo). This measure was chosen as it does not require adjustment for any inflation at KS2, has emerged from our comparisons (using a variety of criteria) in Section 2 as consistently effective at prediction, and the regression framework permits a simple way to incorporate school type. Predictions for 2013, based on 2012 relationships, have been used as the example.

In common with the analysis for Section 2, the dataset was restricted (as described in Section 1.4.2) to results for GCSEs taken in the summer series by 16 year olds for whom a complete set of KS2-based predictions was available. However, by contrast in this section candidates from *all* centre types have been included.

The centre type variable is that used in the National Pupil Database (KS4_NEW_TYPE), which is distributed among our candidates as shown in Table 7.1. Note that there are no candidates from Sixth Form or FE colleges, because the data have already been filtered to include Year 11 candidates only. Academies are not identified as a separate category⁶⁵. Comprehensive centres dominate both in terms of candidates and numbers of entries and candidates from selective centres make up a disproportionately high share of the total entry.

Table 7.1: Distribution of centre type among candidates and entries in 2012 and 2013

Code	Centre type	Percentage of candidates		Percentage of entries	
		2012	2013	2012	2013
0	Invalid	0.0	0.0	0.0	0.0
1	Comprehensive	87.5	87.9	86.6	87.5
2	Selective	4.0	3.9	5.3	5.1
3	Modern	3.7	3.6	3.3	3.2
4	Other Maintained	0.5	0.6	0.3	0.3
5	Independent	4.3	4.0	4.4	3.9
6	Sixth Form College	—	—	—	—
7	Other FE College	—	—	—	—

Two questions explored below are:

- Is the separate treatment of candidates from selective and independent schools justified?
- What would be the result of accounting for them in the modelling rather than excluding them?

⁶⁵ This is beneficial for this analysis, because there have been significant changes in the number and type of academies over the period we are considering. Also, most schools would not have had academy status for the whole of the period during which pupils were studying there.

In answering these, it is important to compare like with like. As such, comparisons with the current method must exclude candidates from selective or independent centres, even if the model is generated using all candidates' results for the base year. It is necessary to consider three models, each fitted using data from candidates taking GCSEs in 2012, as shown in Table 7.2.

Table 7.2: Centre type model

Model ref	Based on candidates	Dependent variables	Notes
C1	Excluding candidates from selective or independent centres	Total normalised marks (K2NrTo)	As fitted in Section 2
C2	All candidates (with restrictions ⁶⁶)	Total normalised marks (K2NrTo)	
C3	All candidates (with restrictions)	Total normalised marks (K2NrTo); centre type	

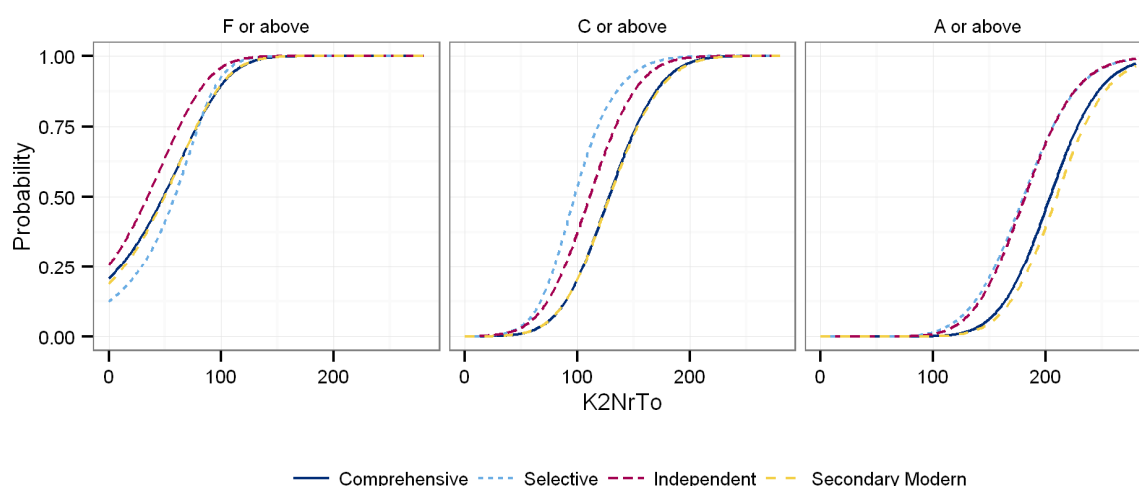
Model C3 was based on that described in Section 2.2 with the addition of dummy variables representing each centre type. We denote these as c_0 – c_5 (with numbering in line with that in Table 7.1). Comprehensive centres have been used as the reference category, and hence a term for c_1 does not appear in the model. The probability that a candidate with a given raw total KS2 mark m would achieve GCSE grade i (where $0=U$, $1=G$, ..., $7=A$, $8=A^*$) in a particular subject is therefore modelled as:

$$\log\left(\frac{p_i}{p_8}\right) = \beta_{0i} + \beta_{1i}m + \beta_{2i}c_0 + \beta_{3i}c_2 + \beta_{4i}c_3 + \beta_{5i}c_4 + \beta_{6i}c_5 \quad (0 \leq i \leq 7)$$

Although there are no interaction terms between centre type and KS2 mark, the model for each grade can have different coefficient values, which allows for the effect of school type to vary over the ability range. Figure 7.1 shows the predictions implied by the model for Mathematics: note that the relative positions of the lines corresponding to candidates from selective and independent centres vary over the grades. No estimate of the statistical significance of the centre type can be made from this model, because it does not account for the inherent multilevel structure of the data (students are grouped within centres). However, the models have been used to generate predictions for 2013, so the extent of any practical differences can be investigated.

⁶⁶ The remaining restrictions other than centre type are still applied to models C2 and C3: Year 11 pupils taking exams in the summer session, with no missing KS2 predictors.

Figure 7.1: Example of predictions under model C3 for Mathematics, using 2012 data



7.1.1 Is the separate treatment of candidates from selective and independent schools justified?

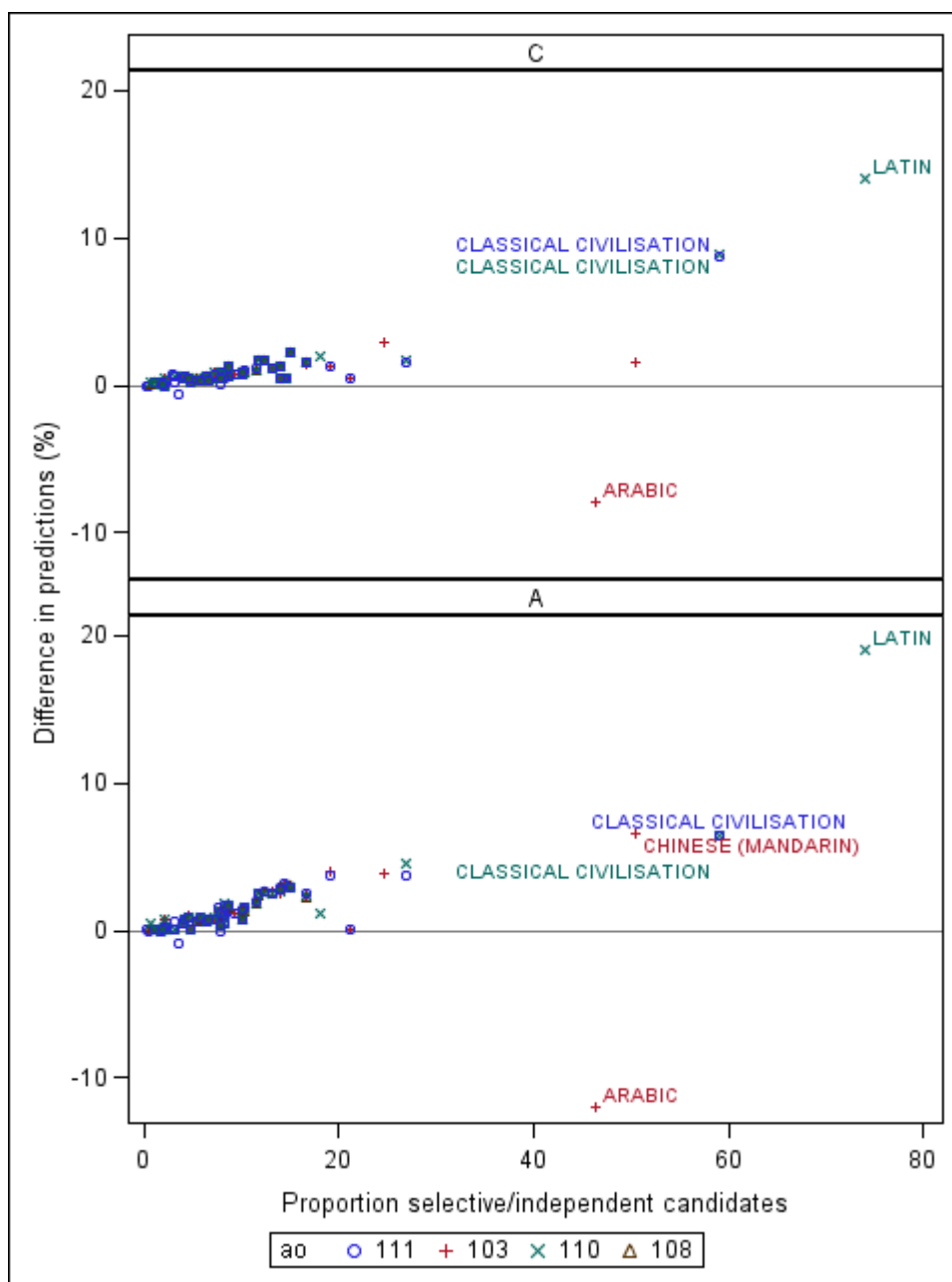
The example in Figure 7.1 gives a hint that the value-added relationship may be different in selective and independent schools, and hence it may be advisable to consider candidates from these centres separately. In order to investigate this more fully across all subjects, predictions for each subject and AO from models C1 and C2 were compared, excluding candidates from selective and independent centres (even though all candidates had been used from the reference year to fit model C2). Table 7.3 shows a summary of the distribution of the difference between predictions at each of the judgemental grades. At grade F the two predictions are virtually indistinguishable, but for the higher grades there is a non-negligible difference, almost always positive: that is, including selective and independent candidates in the model gives more lenient predictions.

Table 7.3 Distribution of difference in predictions (percentage points) by excluding selective/independent candidates from the model (C2 vs C1), excluding subjects with entry less than 400

<i>Grade</i>	<i>Mean</i>	<i>Lower quartile</i>	<i>Median</i>	<i>Upper quartile</i>
F	0.08	0.02	0.06	0.10
C	0.86	0.41	0.58	1.09
A	1.43	0.52	1.10	2.32

Figure 7.2 illustrates the underlying subject data at grades C and A, and shows that the difference in predictions at A is related to the proportion of the entry from selective and independent schools. This effect is not just confined to subjects such as Latin and Classical Civilisation (where over half of the entry comes from selective and independent centres): among the cluster of subjects with around 15 per cent of candidates from selective and independent centres, the difference in predictions is around 2–3 percentage points at grade A.

Figure 7.2: Differences in predictions by excluding selective/independent candidates from the model (model C2 vs C1)



Thus, the value-added relationship is sufficiently different among candidates from selective and independent centres to have an observable impact on predictions, and it is therefore important that it is taken into account in some way.

7.1.2 Does accounting for centre type in the model give any benefit?

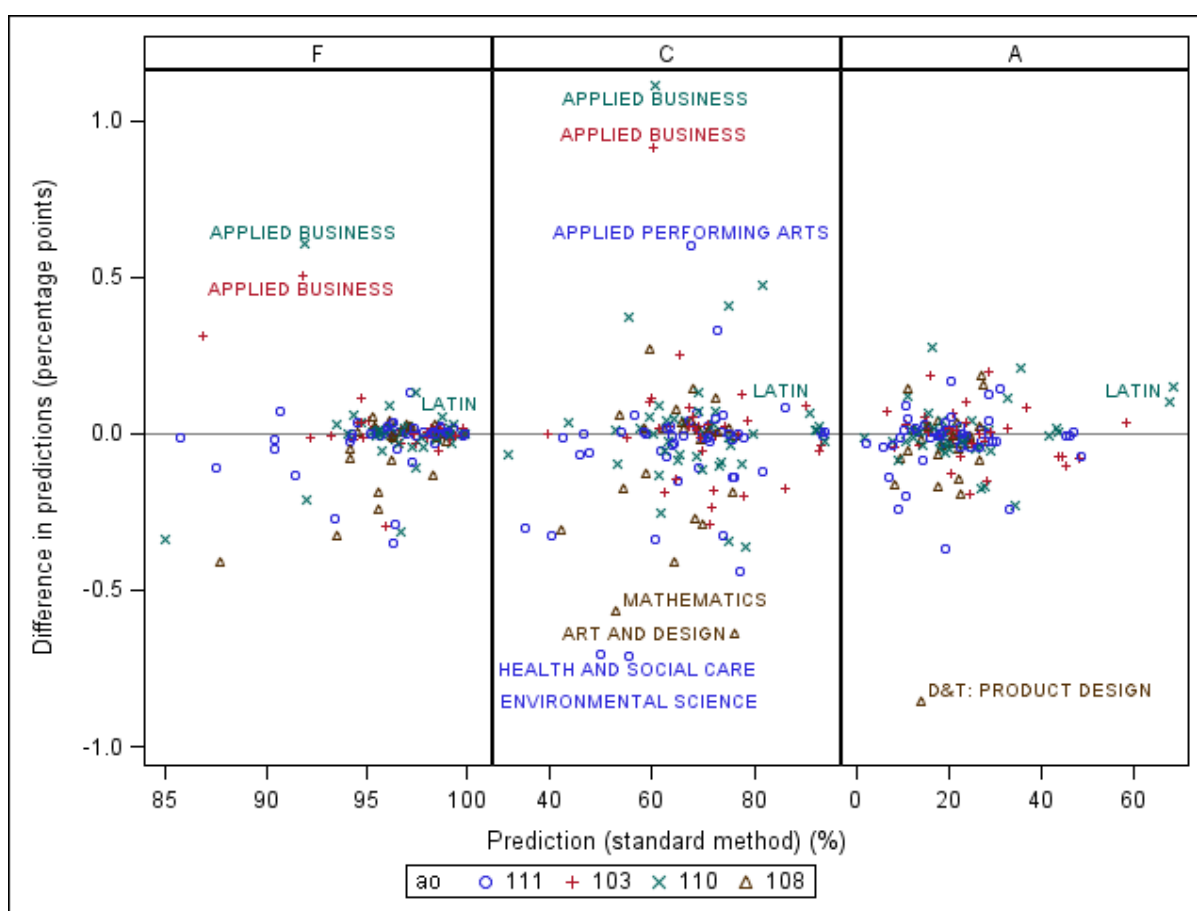
Predictions were generated for each subject and AO from models C3 and C1, excluding candidates from selective and independent centres (even though all candidates had been used from the reference year to fit model C3). Table 7.4 shows a summary of the difference between predictions, where a positive value indicates that model C3 gives a more lenient prediction than C1 for the given set of candidates. At all three grades shown, the differences are small enough to be negligible and they are centred around zero, so there is no evident bias towards leniency or severity.

Table 7.4 Distribution of differences in predictions (at subject/AO level) excluding subjects with entry less than 400 – percentage points

Grade	Mean	Lower quartile	Median	Upper quartile
F	-0.01	-0.01	0.00	0.02
C	-0.02	-0.07	0.00	0.04
A	-0.02	-0.04	-0.01	0.02

Subject/AO combinations where there is a difference of more than 0.5 percentage points have been labelled. There are very few such combinations, and they are almost all within 1 percentage point. Latin has also been labelled, due to the very high proportion of selective/independent candidates, although in fact its differences were relatively small.

Figure 7.3 Differences in predictions by including centre type in model (C3 vs C1, excluding candidates at selective/independent centres)



This has shown that explicitly accounting for centre type gives no benefit, in terms of the predictions actually generated for the candidates not attending selective or independent centres. However, we would expect that the inclusion of more candidates would lead to more accurate predictions for the whole cohort, especially for subjects with a high entry from independent or selective centres (such as Latin). It is not possible to investigate this with the data we have available: raw mark data would be needed to establish cut-scores associated with a given prediction, so that independent and selective candidates could subsequently be slotted in.

The predictions resulting from a model including centre type would be more flexible, especially for subjects with a high independent or selective entry, and at face value it seems more equitable that all candidates are explicitly considered in the model. However, the models using centre type could plausibly be affected by circumstances in just one centre (for less common

centre types and when there are changes in centre type distribution between years). Given that the current framework does allow for changes in the proportion of candidates at selective/independent schools between years, and uses the relationships evident among the majority of the cohort (88%) to indicate the standard and guide the awarding process, with the remaining candidates ‘slotted in’ during grading, there is no compelling reason to move to a more complex model. However, doing so may give a small improvement in prediction and it is something that could be considered in due course, particularly if any further differences emerge in value-added relationships, for example, in academies.

7.2 Controlling for mean centre-level KS2

In Section 4.3 we proposed the use of average level of KS2 achievement within each centre as a possible solution to the problem of KS2 under-predicting inter-board differences.

In this section, we consider the effect on deviance of including this as an additional variable in the logistic regression model based on total normalised scores (K2NrTo). It is calculated at an individual subject/AO level. For example, all pupils in a particular centre entering AQA Geography GCSE. For centres with fewer than five candidates, a mean for candidates across all centres entering the subject with the relevant AO is substituted instead. As with our analysis in Section 7.1, we consider predictions for 2013 using 2012 as a reference year.

Table 7.5 summarises the distribution of subject-level difference in deviance across subjects and shows that the model would reduce deviance by approximately 0.3% on average compared to using logistic regression with normalised score (K2NrTo) alone. For consistency with the analysis in Section 2.5, subjects have been excluded if the full set of 21 predictor variables is not available, or if the entry is less than 400. This extra improvement in predictive power is fairly small compared to the 1.73% reduction in deviance obtained by moving from the current model to using K2NrTo for prediction. The two changes combined would bring a median improvement of 1.89% in subject-level deviance over the current method using KS2 average level.

Table 7.5: Distribution of change in subject-level deviance through inclusion of centre mean KS2, excluding subjects with entry less than 400 or with missing KS2 predictors

Comparator	Change in deviance (%)			
	Mean	Lower quartile	Median	Upper quartile
K2NrTo model	-0.29	-0.53	-0.28	-0.13
K2LevG model	-2.12	-2.94	-1.89	-1.41

7.3 Summary

The analysis in this section has shown that:

- It is important that differences evident in the value-added relationship between different centre types are recognised.
- The current approach to this, by removing selective and independent schools before making predictions, helps to improve accuracy, and there would be very little gain in predictive power from instead attempting to additionally account for school type by statistical modelling.
- Controlling for mean KS2 achievement leads to almost no improvement in the precision of predictions. However, as discussed in Section 4.3, this adjustment may still be worthwhile as it removes the inherent bias in the model and allows the tolerances calculated in Section 3 to be applied more confidently.

8. Further work and final thoughts

8.1 Summary of results

This report has explored a number of the issues relating to the use of KS2-based predictions to set GCSE grade boundaries. For the majority of GCSE awards, from the various analyses carried out in this work to evaluate and improve the generation of predictions, no evidence has emerged to suggest there is anything inappropriate in the current methodology. In general, the evidence in this report is supportive of the way in which KS2 data is used in that:

- Predictions based on KS2 levels are extremely similar to those based upon more detailed ways of quantifying KS2 attainment such as sub-levels, raw scores and normalised scores. Furthermore, analysis has shown that there is very little gain in the predictive power of these models from switching to a new KS2 measure⁶⁷.
- The loss of KS2 Science is likely to have only a very minor impact on the predictive power of models. Furthermore, there are only small differences between predictions constructed including KS2 Science and those constructed without this.
- At grade A, the currently recommended tolerances are roughly in line with those suggested by detailed statistical analysis.
- Even though the correlation is weaker, KS2-based predictions of how far each AO's candidates should be from the national level of attainment in each GCSE subject are very similar to predictions based upon concurrent attainment.
- The current practice of removing selective and independent schools before making predictions helps to improve accuracy, and there would be very little gain in predictive power from instead attempting to additionally account for school type by statistical modelling.

Having said all of the above, there are also some minor areas where the current process for creating GCSE predictions could be improved upon:

- The accuracy of these predictions could be very slightly improved by using logistic regression based upon normalised scores. This is particularly evident in subjects generally taken by high ability candidates such as Separate Sciences. This would also have further advantages as discussed below.
- The current KS2 grade inflation adjustment suffers from not taking account of differences in the prior attainment distribution of candidates in different subjects and in different AOs. This weakness could be addressed through amendments to the calculations but would be automatically addressed if calculations were based on logistic regression using normalised scores or another method not dependent upon the comparability of KS2 levels over time.
- KS2-based predictions tend to under-predict the likely extent of differences between AOs. This becomes particularly evident when predictions are compared to those based on concurrent attainment. The best approach to address this would be to use historical data on the differences between KS2 and screening predictions to adjust future predictions. Where historical data is unavailable or considered unreliable⁶⁸, the best approach is to apply an extension of the use of logistic regression to account, not only for the normalised KS2 scores of individuals, but also the average KS2 attainment level of candidates entering the same subject within their centre.
- Our analysis suggests that the currently recommended tolerances are too low at grade C. If a system was ever developed whereby predictions based upon concurrent attainment could be used during live awarding, then lower tolerances may be applicable.

⁶⁷ Although we do recommend doing this for other reasons detailed below.

⁶⁸ Perhaps due to widespread changes to qualifications.

For the reasons listed above we would recommend that the predictions are no longer based upon KS2 levels and are instead created using logistic regression. Implementing this procedure would require three steps:

1. - Normalised KS2 scores would need to be created based upon national achievement in these tests each year⁶⁹. This task could be completed well in advance of awarding.
2. - A logistic regression model would need to be run for each GCSE subject detailing how the probability of achieving different grades changes according to the prior attainment of candidates. In our experience for this project this has been a relatively straightforward process that could be applied in any standard statistical package⁷⁰.
3. - Once awarding bodies have matched normalised scores to their own GCSE data, this model would then need to be applied to each specification to produce a prediction of the likely percentage of candidates to achieve each grade.

Whether or not this new approach to producing predictions is adopted, we would recommend that the approach to setting tolerances for predictions is made more specific to each award. A simple formula to achieve this is given in Section 3.2.

Although we are recommending that the guideline “tolerances” for GCSEs should be increased, especially at grade C, we have not explored how inter-board comparability would be strongly maintained in this context. Specifically it is not clear how Ofqual could ensure that AOs apply consistent decision processes within this context and how any appearance of a ‘race to the bottom’ could be avoided within the tolerance levels recommended by this report. This issue will require for ongoing discussions between Ofqual and the AOs.

8.2 Other issues not explored

Although this report has attempted to provide a thorough review of the current process for producing GCSE predictions there are remaining issues that have not been touched upon.

One issue for consideration is the extent to which the process for producing GCSE predictions should be transparent and, more importantly, reproducible. At first glance the current process gives the impression of being easily reproducible. Indeed, the very form of the prediction matrices used within the process have a striking similarity to the National Transition Matrices published by the Fischer Family Trust and available to schools through the ‘RAISEonline system’. Furthermore, data on the KS2 prior attainment of candidates entering different subjects with different AOs is widely available to educational researchers via the National Pupil Database. This creates the impression that anyone can easily reproduce the predictions and verify for themselves whether AOs are in line with comparable outcomes. However, this impression is misleading. For example, knowing how AOs have treated early entry, multiple entry and differences between January and June entry is crucial to ensuring that results are correctly reproduced as will making sure that both sets of calculations are based upon the same reference years. Furthermore, it should be noted that, rightly or wrongly, the National Transition Matrices make no adjustment for KS2 grade inflation and so provide different estimates from those calculated by the AOs for the purposes of awarding. Making the process truly transparent and reproducible would require much more detailed, publicly available documentation about the calculations underpinning the process of creating predictions.

In addition to this some of the more fundamental issues regarding how the use of prediction matrices fits with the current accountability system have not been explored. At the heart of this is a fundamental dilemma between whether grades should be awarded to ensure comparable

⁶⁹ In the short term the effect of the KS2 boycott of 2010 would create a particular challenge that would need to be addressed. However, this issue would also create challenges for any method of prediction based upon KS2 results and is not unique to the desire to use normalised scores.

⁷⁰ Furthermore, there were no issues with model convergence for any of the subjects analysed implying that this technique is unlikely to create practical difficulties.

outcomes or to reward comparable performance across years. If GCSE grade boundaries are set purely using predictions based on historical data, then there is nothing in the statistical process that allows genuine improvements in performance to be recognised. There is no statistical mechanism by which this issue can be addressed without further data being used in the process of awarding, such as information from a potential national monitoring test. An alternative would be to allow expert judgment to play some role in determining national standards whilst using statistical information to control inter-board differences. In either case, further research is clearly necessary in order to specify how such a system could work and is beyond the scope of this report.

The effect of a number of particular events on the effectiveness of KS2-based predictions has not been considered within this report. For example, the potential impact of GCSEs becoming linear⁷¹ with the concomitant removal of the January session for GCSE examinations is likely to affect the nature of the candidates taking GCSEs in summer 2014, compared to previous years. As such it will have implications for the ways in which predictions are calculated and applied. On a different note, the widespread boycott of KS2 tests in 2010 will have some impact upon the population of candidates for whom matched data is available in summer 2015 and their comparability to previous years. The effect of these events upon the effectiveness of prediction matrices is very difficult to judge empirically before the cohorts of candidates in question have taken their GCSEs. Rather than address these important but very particular issues, this report has attempted to address some of the overarching issues that will be relevant in every single year when KS2-based predictions are used. The effect of such individual events is something that must be considered and responded to on a case-by-case basis by Ofqual and the awarding bodies.

Similarly, although this report has suggested that the current process is relatively robust, it cannot assess whether the procedure will remain robust in future years as changes to GCSEs are implemented. In particular, it is difficult to assess the likely impact of changing strategies regarding early entry, multiple entry and use of different qualifications (such as IGCSEs and BTECs) on the future accuracy of the method. Furthermore, the introduction of new performance targets for schools could lead to large changes in the number and nature of candidates taking different GCSE subjects. This could lead to a decrease in the level of accuracy of statistical predictions (both from KS2 *and* concurrent attainment) based upon historical performance patterns. This issue will require on-going monitoring as changes to the examination and accountability system take effect.

8.3 Final note

At present the mechanism by which GCSE predictions are produced is relatively straightforward. The only inputs are a simple matrix showing the probability of candidates with differing levels of prior attainment achieving each grade and some information on the numbers of candidates in each prior attainment category. Once these pieces of information are provided, the calculations are sufficiently straightforward that they can be completed on a simple pocket calculator. Whilst the principle of 'keeping things simple' may be very appealing, and whilst this report has provided evidence that is broadly supportive of this approach, the simple nature of the calculations does lead to some weaknesses, albeit relatively minor ones. Specifically, the simple process used at present does not address the issue of KS2 grade inflation fully effectively, tends to slightly under-predict the likely extent of differences in results between AOs and is noticeably, if only slightly, less accurate than a more complex approach for subjects with large numbers of high ability candidates. All of these issues could be addressed more effectively if logistic regression using normalised KS2 scores was used to create predictions. The main criticism of this new approach would be that it does not appear as simple as the approach used currently. However, this criticism would fail to recognise both the technical expertise and the computational

⁷¹ That is, all assessments being completed at the end of two years of study rather than being taken as a series of *units* throughout the course of GCSE.

power available to AOs. In this context, a more complex approach is both well within the capabilities of AOs and could pave the way for further improvements to the use of statistical data and modelling in future.

References

Association of School and College Leaders [ASCL] (2011) *ASCL Technical Guide for Value-Added and other progress measures in RAISEonline and Performance Tables*. Leicester: ASCL.

Benton, T., Hutchison, D., Schagen, I., and Scott, E. (2003) *Study of the Performance of Maintained Secondary Schools in England*. London: National Audit Office.

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

Benton, T., and Sutch, T. (2012) *Exploring the value of GCSE prediction matrices based upon attainment at Key Stage 2*. Cambridge Assessment Internal Report.

Department for Education [DfE] (2011) *National Pupil Database: KS4 User Guide 2011*. Available at: <http://www.bristol.ac.uk/cmipo/plug/support-docs/ks4userguide2011.pdf> (Accessed on 3 December 2013). Eason, S. (2003) *AQA GCSE Science - Summer 2002: Centres Splitting Entries Across Options*. Research report RPA_03_SE_RP_017. Guildford: AQA.

Eason, S. (2006) *Stability of Common Centres' Analyses*. Research report RPA_06_SE_WP_035. Guildford: AQA.

Eason, S. (2009) *Predicting GCSE outcomes based on past centre-level performance data*. Research report RPA_09_SE_TR_008. Guildford: AQA.

Eason, S. (2010) *Predicting GCSE outcomes based on candidates' prior achieved key stage 2 results*. Research report RPA_10_SE_TR_001. Guildford: AQA.

Eason, S. (2012) *Alternative key stage 2 models for predicting GCSE outcomes*. Research report CERP_TR_SE_13092012. Guildford: AQA.

Hutchison, D. (2007) When is a compositional effect not a compositional effect?, *Quality and Quantity*, 41, 219-232.

Newton, P. (2011) A level pass rates and the enduring myth of norm-referencing. *Research Matters: A Cambridge Assessment Publication*, Special Issue 2, 20-26.

Ofqual (2012) *GCSEs and A Levels in Summer 2012: Our approach to maintaining standards*. Available at: <http://ofqual.gov.uk/documents/gcses-and-a-levels-in-summer-2012-our-approach-to-setting-and-maintaining-standards/> (Accessed on 3 December 2013).

Ofqual (2013) *Summer 2013 Data Exchange procedures: GCE, GCSE and Level 1 / 2 certificates*. Available at: <http://ofqual.gov.uk/documents/summer-2013-data-exchange-procedures/> (Accessed on 3 December 2013).

Pinot de Moira, A. (2008) *Statistical Predictions in Award Meetings: How confident should we be?* Research report RPA_08_APM_RP_013. Guildford: AQA.

Ridgeway, G. (2012) *Generalized Boosted Models: A guide to the gbm package*. Available at <http://gradientboostedmodels.googlecode.com/git/gbm/inst/doc/gbm.pdf> (Accessed on 4 December 2013).

Smith, A. (2013) *The effectiveness of Key Stage 2 results as a predictor of GCSE attainment*. Cardiff: WJEC.

Spalding, V. (2012) *Predicting GCSE outcomes based on candidates' concurrent GCSE English and Maths results*. Unpublished research report.

Treadaway, M. (2013) *An analysis of Key Stage 2 reliability and validity*. FFT Research Paper No. 2. Cowbridge: FFT Education Limited.

Wang, T. and Brennan, R. (2009) A Modified Frequency Estimation Equating Method for the Common-Item Nonequivalent Groups Design. *Applied Psychological Measurement*, 33, 118-132.

Appendix 1: GCSE predictions using mean Key Stage 2 Level as the measure of prior attainment

Provided by Martin Taylor, AQA, October 2013

This document describes how to calculate predictions for the current year y from the aggregated outcomes in reference years x_1 and x .

1. - Candidates for the GCSE in question are allocated to one of eight categories according to their mean Key Stage 2 levels.

Candidates are included in predictions only if they have three Key Stage 2 results in the relevant year⁷². A similar comment applies to the candidates in the reference year(s), ie they must have three KS2 levels. These are called *matched candidates*.

2. - The mean Key Stage 2 cut-offs are as follows.

Category	Cut-off
8	4.668
7	4.334
6	4.001
5	3.668
4	3.334
3	3.001
2	2.668
1	0

For example, category 7 includes mean KS2 levels between 4.334 and 4.667 inclusive. In fact, mean KS2 levels can be only exact thirds, ie 0, 0.333, 0.667, 1.000, 1.333, etc. Therefore, in practice category 7 contains just 4.667.

3. - Produce outcome matrices for the subject in question for years x_1 and x (see Table 1).

⁷² 'Relevant year' is the year in which they were age 11 and therefore did KS2 tests. For example if GCSE predictions for 16 year-olds in 2013 are being determined, the relevant year is 2008 (five years previously).

Table 1 Example of outcome matrices for years x_1 and x_2

(a) Outcome matrix for year x_1

Category	No. of cand.s.	A*	A	B
1	1235	41.42
2	2146	20.24
3	1815	8.59
.
.
8
Total	7915

(b) Outcome matrix for year x_2

Category	No. of cand.s.	A*	A	B
1	1016	43.49
2	1995	23.42
3	1347	8.02
.
.
8
Total	7647

4. - Produce two sets of raw predictions for year y , one based on year x_1 and one based on year x_2 (see Table 2). (These are called raw predictions because they take no account of Key Stage 2 inflation/deflation.) To do this, use the outcome matrix from Table 1 with the year y entries in the 'No. of cand.s' column.

Table 2 Predictions for year y

(a) Based on year x_1

Category	No. of cand.s.	A*	A	B
1	945	41.42
2	1357	20.24
3	1068	8.59
.
.
8
Total	6416	17.45

The raw prediction at grade A is

$$(945 \times 41.42 + 1357 \times 20.24 + 1068 \times 8.59 + \dots) \div (945 + 1357 + 1068 + \dots)$$

$$= 17.45 \text{ (say).}$$

(b) Based on year x

Category	No. of cand.	A*	A	B
1	945	43.49
2	1357	23.42
3	1068	8.02
.
.
8
Total	6416	17.96

The raw prediction at grade A is

$$(945 \times 43.49 + 1357 \times 23.42 + 1068 \times 8.02 + \dots) \div (945 + 1357 + 1068 + \dots)$$

$$= 17.96 \text{ (say)}$$

5. - Because of the discrete nature of the mean KS2 levels, it is not possible to control for KS2 inflation by adjusting the cut-offs. Instead, the cut-offs remain fixed and the adjustments are the changes which would occur (between the reference year and the current year) if the whole national cohort of candidates was entered for the GCSE subject in question. (The inflation adjustments are therefore subject-specific but not specification-specific.)

Therefore, produce predictions for year x_1 for the subject in question if it is assumed that *all* pupils who have three KS2 levels in the relevant year (five years prior to x_1 , if predictions for 16 year-olds are being calculated) entered for the GCSE subject in question in year x_1 (see Table 3(a)). These predictions are based on the actual outcomes in year x_1 (see Table 1(a)).

In a similar way, produce predictions for years x and y for the GCSE subject in question if all pupils had taken it. The predictions for year x (Table 3(b)) are based on the actual outcomes in year x (see Table 1(b)). There are two sets of predictions for year y (Tables 3(c) and 3(d)): one set based on the actual outcomes in year x_1 and another set based on the actual outcomes in year x (see Tables 1(a) and 1(b)).

Table 3 Predictions for the GCSE subject in question if all pupils with three KS2 results in the relevant year had taken it

(a) Predictions for year x_1 if all pupils with these KS2 results in year $(x_1 - 5)$ had taken the GCSE

Category	No. of cand.	A*	A	B
1	105076	41.42
2	93504	20.24
3	95779	8.59
.
.
8
Total	596577	8.72

The prediction at grade A is

$$(105076 \times 41.42 + 93504 \times 20.24 + 95779 \times 8.59 + \dots) \div (105076 + 93504 + 95779 + \dots) \\ = 8.72 \text{ (say).}$$

Note that the percentages in the body of the table come from Table 1(a).

(b) Predictions for year x if all pupils with three KS2 results in year $(x - 5)$ had taken the GCSE

Category	No. of cand.	A*	A	B
1	104689	43.49
2	94706	23.42
3	96812	8.02
.
.
8
Total	602675	9.24

The prediction at grade A is

$$(104689 \times 43.49 + 94706 \times 23.42 + 96812 \times 8.02 + \dots) \div (104689 + 94706 + 96812 + \dots) \\ = 9.24 \text{ (say).}$$

Note that the percentages in the body of the table come from Table 1(b).

(c) Predictions for year y , based on year x_1 , if all pupils with three KS2 results in year $(y - 5)$ had taken the GCSE

Category	No. of cand.	A*	A	B
1	109826	41.42
2	97835	20.24
3	96702	8.59
.
.
8
Total	609884	8.94

The prediction at grade A is

$$(109826 \times 41.42 + 97835 \times 20.24 + 96702 \times 8.59 + \dots) \div (109826 + 97835 + 96702 + \dots)$$

$$= 8.94 \text{ (say).}$$

Note that the percentages in the body of the table come from Table 1(a).

(d) - Predictions for year y , based on year x , if all pupils with three KS2 results in year $(y - 5)$ had taken the GCSE

Category	No. of cand.	A*	A	B
1	109826	43.49
2	97835	23.42
3	96702	8.02
.
.
8
Total	609884	9.51

The prediction at grade A is

$$(109826 \times 43.49 + 97835 \times 23.42 + 96702 \times 8.02 + \dots) \div (109826 + 97835 + 96702 + \dots)$$

$$= 9.51 \text{ (say).}$$

Note that the percentages in the body of the table come from Table 1(b) and that the numbers of candidates are the same as in Table 3(c).

Thus the predictions at grade A for years x_1 and x , if all pupils with three KS2 results in the relevant years took the GCSE subject in question, are 8.72 and 9.24 (Tables 3(a) and 3(b)). The similar predictions at grade A for year y , based on years x_1 and x respectively, are 8.94 and 9.51 (Tables 3(c) and 3(d)).

6. Calculate adjusted predictions for year y based on years x_1 and x .

The adjusted prediction based on year x_1 is

(raw prediction from Table 2(a) + all-candidates prediction for year x_1 from Table 3(a) - all-candidates prediction for year y from Table 3(c)).

At grade A this is

$$17.45 + 8.72 - 8.94 = 17.23.$$

The adjusted grade A prediction based on year x is calculated in a similar way, ie

$$17.96 + 9.24 - 9.51 = 17.69 -$$

(see Tables 2(b), 3(b) and 3(d)). -

7. Calculate an (adjusted) prediction based on years x_1 and x combined.

This is simply the weighted mean of the two predictions taking account of the total entry in the reference years.

At grade A this is -

$$(7915 \times 17.23 + 7647 \times 17.69) \div (7915 + 7647) = 17.45 -$$

(figures taken from Table 1 and from section 6). -

Appendix 2: Detailed description of methodology used to estimate tolerances for each AO and each subject

This section describes the methodology used to estimate the standard errors around the predicted percentages of students to achieve each GCSE grade or above. The process used to calculate standard errors was Fay's method of balanced repeated replication. The crucial advantages of this technique are that it accounts for variability in the effects of different centres, as well variability in performance between candidates within these centres, and that it requires very few assumptions.

The basic idea of the method is to repeatedly recalculate the quantities we are interested in based upon a randomly chosen half of the available centres within the data. The extent of variation between different half samples is related to the standard error of the quantity we are interested in by a known formula. In other words if we get very different answers when we recalculate the quantity of interest with different half samples we know that there is a large standard error. If different half samples give very similar results then the standard error must be small.

The method was applied to each GCSE subject in turn as follows:

1. - Create a list of centres entering candidates for the subject in any of June 2011, June 2012 and June 2013.
2. - Split all of these centres into 56⁷³ strata based upon the total number of GCSE entries in this subject within centres across the three years.
3. - Within each strata, sort the centres by the AO with which they have entered their candidates⁷⁴, the number of years in which they entered candidates for the subject and the number of pupils they have entered for the subject in total.
4. - Now assign each centre in each strata to one of two variance PSUs (primary sampling units). This is done by assigning each successive centre in our sorted list within each strata to an alternate PSU. In plain language this means that, within each strata, we have split the centres into two groups. The sorting in stage 3 helps to ensure a stable distribution of centres across AOs and years as well as a stable total entry size between the two groups.
5. - Now begin randomly sampling half of the centres within the data. This is done using a Hadamard Matrix of size 56⁷⁵. Fifty-six sets of weights are now generated dependent upon the 56 rows of the Hadamard matrix:
 - a. - If the *j*th number in that row is equal to 1 then all of the centres within the first variance PSU within the *j*th strata are given a weight of 1.5 and the centres in the second variance PSU are given a weight of 0.5.
 - b. - If the *j*th number in that row is equal to -1 then all of the centres within the first variance PSU within the *j*th strata are given a weight of 0.5 and the centres in the second variance PSU are given a weight of 1.5.
 - c. - These weights are now adjusted so that, for each AO, the distribution of prior attainment at KS2 of candidates taking their GCSEs in 2013 is kept equal to the overall distribution for this AO in 2013 in the original (unweighted) data.
6. - For each of the fifty-six sets of weights in turn we now calculate:

⁷³ Previous work to examine standard errors in A level predictions (Benton and Lin, 2011) used 80 strata. The decision to use 56 strata was taken in order to reduce the computational burden of calculations. This is unlikely to have a noticeable impact upon any substantive conclusions. Further validation of the approach, demonstrating the robustness of the technique, is given at the end of the appendix.

⁷⁴ With a separate AO code denoting if they have entered candidates with more than one AO across the three years.

⁷⁵ A Hadamard matrix is a square matrix of a given size containing the values 1 and -1. It is specifically designed so that each row is mathematically orthogonal to all the others. The 56 by 56 Hadamard matrix used for analysis was provided via the *survey* package in R written by Thomas Lumley (<http://cran.r-project.org/web/packages/survey/index.html>). Broadly speaking the aim of using Hadamard matrices within this context is that whilst we are randomly selecting half samples we ensure that each of the half samples are sufficiently different from one another (that is, we don't inadvertently always select the same centres in every half sample).

- a. - The predicted percentage to achieve each grade or above with each AO in 2013 based on the prediction matrices from 2011 and 2012 using the method described in Appendix 1.
 - b. - The actual percentage achieving each grade or above with each AO in 2013.
 - c. - The difference between the actual and the predicted percentage achieving each grade or above⁷⁶.
7. - We can now estimate the standard error of the predictions matrices method as a whole for each AO within the subject in turn by Fay's formula:

$$\text{Standard Error} = \sqrt{\frac{\sum(d_j - d)}{56(1 - 0.5)}}$$

Where d is the estimate of the difference between predicted and actual outcomes for the whole sample and d_j is the difference based upon the j th set of weights.

A few caveats should be noted around these standard errors:

1. - The methodology here only examines random error (that is variation between centres) rather than any systematic errors in the method. Thus, using estimates derived in this way to assign tolerances, assumes that the relationship between KS2 attainment and achievement in GCSEs is consistent across years. Essentially we are assuming that the prediction matrices developed in 2011 and 2012 are still trustworthy for 2013 data.
2. - The reliance of the above formula on the difference between actual and predicted grade distributions is only technically correct if (in the overall data for each AO in a subject) the predicted percentage is equal to actual percentage achieving each given grade in each year. Generally speaking these percentages are very close. However, in the small minority of cases where they are very different, the standard errors estimated by the above process will be less reliable.

A similar method to the one described above was used to calculate the standard error of the difference between predictions based upon KS2 and predictions based upon concurrent attainment.

Further validation of the method

In order to check the validity of the method above an alternative methodology was also applied based upon multilevel modelling. The essential idea behind this alternative method was to fit a multilevel model to the data for each subject and then use the results to examine how much variability we would expect between predicted and actual results if the model reflected reality.

The process for each GCSE subject was as follows:

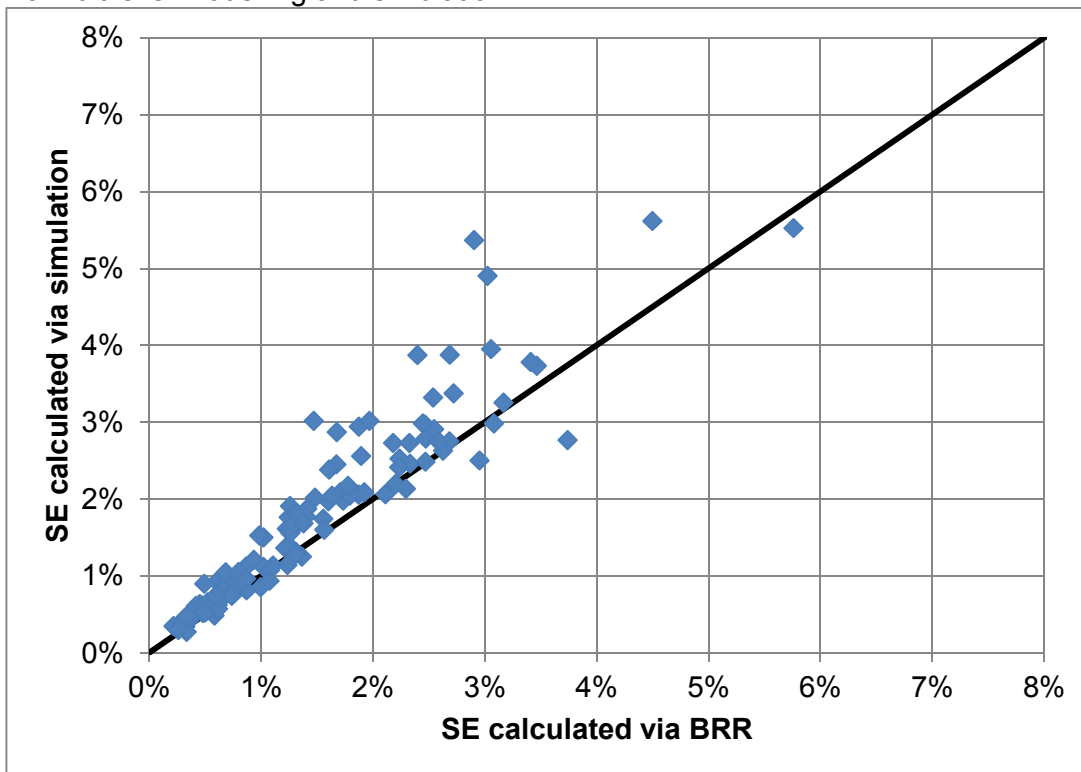
1. - Using all of the available data across all AOs from June 2011, June 2012 and June 2013, fit a multilevel logistic model examining the probability of candidate achieving grade C or above. Specifically this estimates the following:
 - a. - Eight fixed coefficients examining the relationship between the probability of achieving a C or above and the eight categories defined by average KS2 level.
 - b. - A random coefficient that estimates the variance in the effects of different centres on this probability.

⁷⁶ The focus on the *difference* between actual and predicted outcomes at this stage combines the “model standard errors” and “innate standard errors” described by Benton and Lin (2011) into a single step. In theory this may improve the accuracy of the method as it accounts for the fact that it is likely that the same centres remain within a subject across years so that if predictions are higher for a particular half-sample then actual results are also likely to be higher. However, a method closer to that applied by Benton and Lin was applied by Benton and Sutch (2012) and found to give very similar results to the ones given in this report indicating that this change does not have a dramatic impact upon results.

- c. - A random coefficient that estimates the variance of the effect of a centre between years.
2. - Using the coefficients estimated in 1b and 1c simulate plausible effects of each centre in each year.
3. - Using the coefficients estimated in 1a and the effects simulated in step 2 calculate the probability that each candidate in each year will achieve a grade C or above.
4. - Using the probabilities created in step 3, simulate achievement data for each candidate (i.e. whether they achieve grade C or above).
5. - Based on the simulated data for 2011 and 2012, and using the usual predictions matrices technique, generate predictions of the percentage of candidates who will achieve grade C or above in 2013 within each AO.
6. - Calculate the simulated percentage who actually achieve grade C or above in 2013.
7. - Calculate the difference between simulated predicted outcomes (stage 5) and simulated actual outcomes (stage 6) for each AO in 2013.
8. - Repeat steps 2 to 7 two hundred times⁷⁷ and record the difference between predicted and actual outcomes for each AO for each simulation.
9. - The standard deviation of the differences between predicted and actual outcomes across the 200 simulations for each AO now provides an estimate of the standard error of the prediction matrices procedure.

The above approach was completed for all subjects with the exceptions of English, Maths and Environmental Science. English and Maths were excluded due to the computational burden of fitting and applying the necessary multilevel models with the extremely large numbers of entries in these subjects. Environmental Science was excluded due to the difficulties examined earlier with estimating a meaningful standard error when such a large proportion of the candidates are found in a very small number of centres.

Figure A2.1: A comparison of standard errors estimated via BRR with standard errors estimated via multilevel modelling and simulation



⁷⁷ Ideally this technique would use more than 200 simulations for each subject. However, due to the computational burden of the procedure, only a relatively small number of simulations were used.

The standard errors estimated via the above approach are compared to the standard errors estimated via BRR in Figure A2.1. As can be seen, there is a very close level of agreement between the two sets of estimates (correlation=0.94). This provides evidence that the tolerances provided in this report are generally robust to alternative methods of estimation.

Further inspection of Figure A2.1 indicates that the standard errors estimated via simulation tend to be very slightly higher than those estimated via BRR. However, there are good reasons to believe that the estimates from BRR are the more reliable of the two. In particular, further inspection of the individual multilevel models for each subject revealed that all of these showed evidence of under-dispersion. In other words, the results indicated that the degree of variability in the achievement of the population of candidates with a fixed level of prior attainment was less than would be expected given the underlying assumptions of the model. Essentially this indicates a general lack of fit in the multilevel models and, more importantly, suggests that any simulations based upon the model coefficients are likely to overestimate the degree of variability in results. This further endorses the use of BRR as the most appropriate estimation technique.

Appendix 3: A modified method for producing GCSE predictions based upon Key Stage 2

The calculations in this Appendix relate to the discussion in Section 4.3.1 of the report concerning adjusting KS2-based prediction to account for the low correlation between KS2 and GCSE achievement. The aim of such an adjustment is to address the tendency for KS2-based predictions to under-predict the extent of differences between AOs.

We are interested in using data from population 1 (the national population within a particular subject in the reference year) to set GCSE grade boundaries in population 2 (the population for the same subject in the current year for a particular AO). First, we define V_1 and V_2 to be the values of KS2 prior attainment in each population. Next, we define C_1 and C_2 to be the expected value of concurrent attainment in each population on a scale devised so that the expected values and variances of C_1 and C_2 are designed to equal the expected values and variances of V_1 and V_2 . We can then define the expected values of concurrent attainment for any individual given their individual level of prior attainment as follows:

$$\begin{aligned} E(C_1) &= Mean_1(V) + Cor_1(V, C)[V_1 - Mean_1(V)] \\ E(C_2) &= Mean_2(V) + Cor_2(V, C)[V_2 - Mean_2(V)] \end{aligned}$$

Where $Mean_k(V)$ is the average level of prior attainment in population k , and $Cor_k(C, V)$ is the correlation between prior attainment and concurrent attainment in population k . According to the argument of Wang and Brennan (2009) we need to adjust each V_2 so that it is equal to the value of prior attainment in population 1 associated with the same level of expected concurrent attainment. This is done by solving $E(C_1) = E(C_2)$. This implies that we should adjust each V_2 to:

$$Adjusted V_2 = Mean_1(V) + \frac{[Mean_2(V) - Mean_1(V)]}{Cor_1(V, C)} + \left(\frac{Cor_2(V, C)}{Cor_1(V, C)} \right) [V_2 - Mean_2(V)]$$

Predictions should then be based on these adjusted V_2 values rather than the original ones. In order to apply this adjustment, one further modification is required. Because the correlation between KS2 and concurrent GCSE cannot be calculated empirically in population 2 (that is, $Cor_2(V, C)$) until all GCSE awarding is completed, it is necessary to estimate this based upon the correlation found in population 1 (that is, the reference year). This can be done simply by adjusting the correlation calculated in the reference year according to the different standard deviations of prior attainment in the two populations (S_1 and S_2) as follows:

$$Estimated Cor_2(V, C) = \frac{S_2 Cor_1(V, C)}{\sqrt{(S_2^2 Cor_1(V, C)^2 + S_1^2 - S_1^2 Cor_1(V, C)^2)}}$$

Note that even if V_2 is defined to be one of a specific set of values (such as mean KS2 levels, 2.66, 3.00, 3.33, ...) the adjusted V_2 values are unlikely to match these specific values. Wang and Brennan (ibid) suggest that this issue can be overcome by using linear interpolation to estimate the probability of candidates achieving each GCSE grade for different adjusted V_2 values. This additional complication is avoided if the V s (that is the measures of KS2 prior attainment) are defined on a continuous scale such as normalised scores. If such scores are combined with logistic regression, then converting the adjusted KS2 scores into probabilities of candidates achieving each GCSE grade is entirely straightforward.

Appendix 4: Examination of the relationship between KS2 match rate and agreement of results with screening outcomes

The aim of the analysis in this section is to explore the extent to which discrepancies between KS2-based and screening predictions may be caused by low KS2 match rates. The difficulty with this analysis is that each set of predictions is based upon a different population. KS2-based predictions require matching KS2 data but not matching data on concurrent attainment. Screening predictions require matching data on concurrent attainment but not matching data on KS2. This means that it cannot be assumed that discrepancies between the two predictions necessarily indicate a problem with either. Ideally this would be addressed by comparing the grade boundaries suggested by each set of predictions. However, data on raw marks achieved was not available for analysis and so this approach was not possible.

As an alternative, in order to make meaningful comparisons, we have used the actual achievement of candidates for each AO/subject combination as a fixed point for analysis. We can firstly calculate the difference between centred KS2-based predictions⁷⁸ and centred actual achievement⁷⁹ for candidates with matching KS2 data. This gives a KS2-based estimate of the extent to which awarding might be viewed as lenient or severe. We can then also calculate the difference between centred screening predictions and actual achievement for candidates with matching concurrent achievement data. This gives a second, screening-based estimate of the extent to which awarding might be viewed as lenient or severe. Finally, we calculate the difference between these two estimates. A difference between these two estimates of leniency, (for example, if KS2 suggests an award was lenient whilst, for the exact same qualification, screening suggests an award was harsh), may indicate a problem with the predictions from one of these sources.

Analysis is restricted to subjects where at least two AOs have a matched entry of at least 500 candidates. Analysis is based upon GCSEs awarded in June 2013.

The aim of this analysis is to examine the association between differences in the estimated leniency of awards using different sources of data and the KS2 match rate. For this purpose, the KS2 match rate is defined as the percentage of candidates with matching concurrent GCSE data who also have matching KS2 data and are not located in independent or selective schools⁸⁰.

The relationship between KS2 match rate and the absolute difference between screening and KS2-based estimates of leniency at grade C is shown in Figure A4.1. This shows that, where the KS2 match rate was greater than 60 per cent, with a small number of exceptions, there tended to be a good level of agreement between KS2 and screening-based estimates of leniency. However, where the KS2 match rate was below 60 per cent, the two estimates of leniency disagreed more frequently. Specifically, in two-thirds (12 out of 18) of the instances where the KS2 match rate was below 60 per cent, the absolute difference between KS2 and screening-based estimates of the leniency of awards was greater than 1 percentage point⁸¹. Conversely, in about three quarters of instances where the KS2 match rate was above 60 per cent (90 out of 119), KS2 and screening-based estimates of the leniency of awards differed by less than 1 percentage point.

Similar results were found at grade A (Figure A4.2). Again, in more than half (13 out of 18) of the instances where the KS2 match rate was below 60 per cent, the absolute difference between

⁷⁸ As with other analyses of screening predictions in this report, we make use of centred predictions. That is, the extent to which predictions are above or below the national average.

⁷⁹ That is, the extent to which actual percentage of pupils achieving a given grade or above within an AO is above or below the national average.

⁸⁰ That is, their KS2 data is matched and would be used in generating prediction matrices.

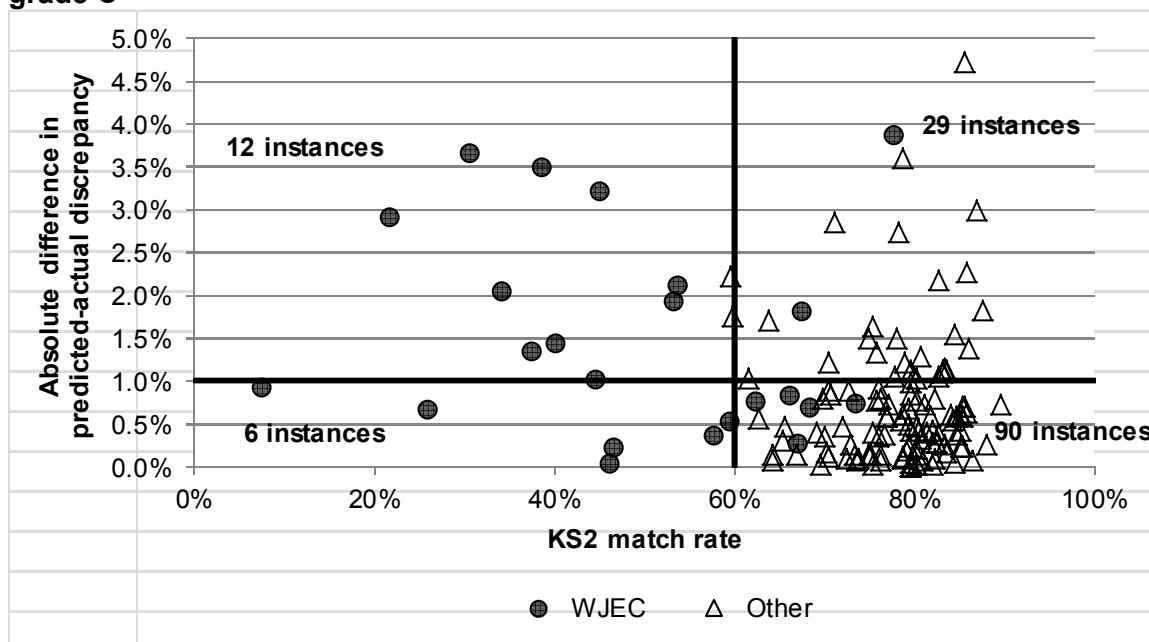
⁸¹ Very roughly, across the range of different match rates, differences of more than 1 percentage point tended to represent statistically significant differences between the two estimates (based upon further calculations using balanced repeated replication).

KS2 and screening-based estimates of the leniency of awards was greater than 1 percentage point. Conversely, in more than three quarters of instances where the KS2 match rate was above 60 per cent (95 out of 119), KS2 and screening-based estimates of the leniency of awards differed by less than 1 percentage point.

This analysis suggests that, where the KS2 match rate falls below 60 per cent there is a greater risk of screening and KS2-based methods providing different results. However, one caveat on this analysis is that, as can be seen in Figure A4.1, almost all⁸² of the AO/subject combinations with low KS2 match rates are provided by WJEC. Furthermore, for these GCSEs, the low match rates are largely the result of Welsh candidates, who have never taken KS2 tests. Thus, not only is the match rate lower for WJEC in these cases, the pupils with matching KS2 data are very clearly non-representative of pupils as a whole. It cannot be determined whether the same results would occur for low match rates caused in different ways.

If we rerun the same analysis but restrict it to English candidates⁸³ before we start, then, across all AOs, there is little variation in match rates. In fact we find that for every AO/subject combination the KS2 match rate is above 60 per cent. Furthermore, within this very restricted range, we cannot identify any clear relationship between KS2 match rates and the size of differences between KS2 and screening-based estimates of the leniency of awards. This implies that KS2 match rates are not a sufficiently strong influence on the accuracy of KS2-based predictions for us to be able to detect their influence over and above other factors that may affect the accuracy of the method.

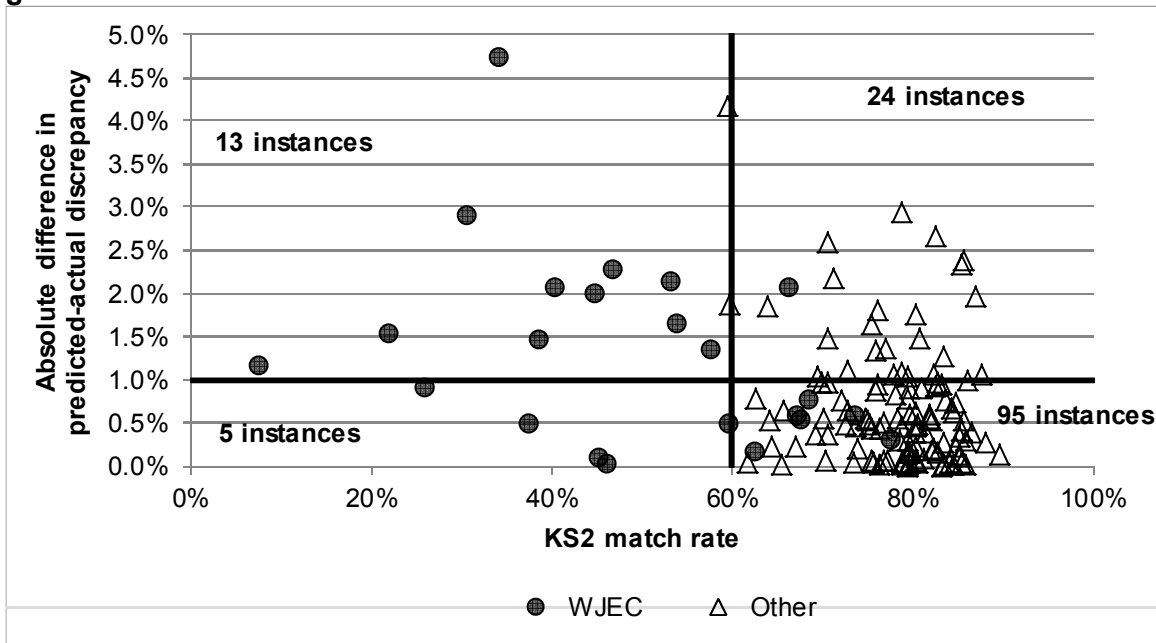
Figure A4.1: The relationship between KS2 match rate and absolute differences in estimates of leniency/severity of awards based upon screening and based upon KS2 at grade C



⁸² With just two exceptions.

⁸³ That is, candidates studying within centres in England.

Figure A4.2: The relationship between KS2 match rate and absolute differences in estimates of leniency/severity of awards based upon screening and based upon KS2 at grade A



We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

Published by the Office of Qualifications and Examinations Regulation in 2014

© Crown copyright 2014

You may re-use this publication (not including logos) free of charge in any format or medium, under the terms of the [Open Government Licence](#). To view this licence, visit [The National Archives](#); or write to the Information Policy Team, The National Archives, Kew, Richmond, Surrey, TW9 4DU; or email: psi@nationalarchives.gsi.gov.uk

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346