CAMBRIDGE ASSESSMENT

Formalising and evaluating the benchmark centres methodology for setting GCSE standards

Tom Benton

Cambridge Assessment Research Report

27th March 2013

University of Cambridge
Local Examinations Syndicate

Author contact details:
Tom Benton
ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
Benton.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk/

How to cite this publication:

Benton, T. (2013). *Formalising and evaluating the benchmark centres methodology for setting GCSE standards*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Table of Contents

Introduction 4
The relative value of data from common centres and key stage 2 5
Centre level correlations analysis 5
Comparison of putative grade distribution with “gold standard” 6
The value of redefining “benchmark” centres 8
Definitions based on absolute entry size 10
Definitions based relative change in entry size 12
Definitions based historical stability of results 14
Definitions based stable prior attainment 17
The optimal definition of a benchmark centre 19
Using more than one year of historical data 21
Summary 23
References 24
Appendix A: Correlations with centre level achievement for each OCR GCSE specification in summer 2012 25
Appendix B: OCR GCSE Specifications used in exploration of different definitions for benchmark centres 26

Introduction

One of the most straightforward ways of predicting likely examination results in any subject this year is simply to look at what they were last year and predict that they will be the same. The use of such easily available statistical information has been often used to inform the awarding process since such data provides a check on whether changes in achievement between years are plausible. A slightly enhanced version of this method is to use *common centres*. A common centre is a centre that has entered students for a subject in two successive years. The assumption is that the centre's results are unlikely to be very different in those two years. On the basis of this assumption one possible approach to maintaining examination standards is to set grade boundaries so that, across common centres, the percentage of pupils achieving any given grade remains as consistent as possible.

Ofqual explicitly encourage that this approach to maintaining standards is applied alongside the more commonly discussed *prediction matrices* approach based on the prior attainment of candidates at key stage 2.

> "…the regulators have agreed with exam boards that emerging results in August 2012 will be reported to the regulators using two measures. *All exam boards will report their outcomes compared to the results achieved by common centres from 2011*. In addition, the three exam boards based in England will report their outcomes against predictions for the cohort based on prior achievement at Key Stage 2."[1] (emphasis added)

However it should be noted that in the case of AQA, Edexcel, and OCR it is clear that it is intended that data from common centres should be given less importance than predictions based upon key stage 2.

> "AQA, Edexcel and OCR should report any out of tolerance outcomes against KS2-based predictions. WJEC and CCEA should report any out of tolerance outcomes against common centre predictions."[2]

In combination, the above quotes show that Ofqual wishes awarding bodies to explicitly report outcomes for all qualifications compared to results in common centres the previous year. However, only WJEC and CCEA are encouraged to keep results within tolerance of these predictions. In contrast AQA, Edexcel and OCR are not required to give any weight to the predictions based on common centres – only to calculate them.

A similar (but not necessarily identical) approach is the use of *benchmark centres*. The precise definition of a benchmark centre has never been formally described. However, the rationale is essentially to identify centres that have entered candidates for a given subject in the past and where we feel especially confident in expecting that their results will be consistent. The exact grounds for such confidence are not clearly defined but may include any of the following:

- Large numbers of pupils entering a given subject historically.
- Stable numbers of pupils entering a given subject in successive years.
- Historically stable results within the given subject.
- Stable pupil background characteristics particularly in terms of their prior attainment.

Depending upon how exactly we define the term "stable" in the above, it is clear that a benchmark centres methodology could range from being identical to the common centres

[1] From Ofqual's "Approach to setting and maintaining standards" downloaded from http://www.ofqual.gov.uk/files/2012-05-09-maintaining-standards-in-summer-2012.pdf on 27th February 2013.
[2] From Ofqual's "Procedures for summer 2012 GCE and GCSE data exchange" downloaded from http://www.ofqual.gov.uk/files/2012-06-28-summer-2012-gce-and-gcse-data-exchange-procedures.pdf on 27th February 2013.

approach recommended by Ofqual to being a similar approach but based upon a much more select group of centres.

The aim of this research report is to explore empirically the value of the common centres approach to maintaining standards. Furthermore, the research will explore the issue of whether there is any value in pursuing a more nuanced version of the common centres approach via a formal definition of the characteristics of benchmark centres. The specific questions addressed by this report are:

- What empirical evidence is there of the value of a common centres approach to maintaining standards compared to one based on key stage 2 based prediction matrices?
- Does restricting benchmark centres to those with historically stable results, large entries or consistent entries help to improve the accuracy of predictions?
- Should future performance of centres be predicted from several years' worth of historical data or only from the most recent performance?

The relative value of data from common centres and key stage 2

Two methods were employed to examine the extent to which data from common centres may be more or less valuable than candidate's prior attainment data from key stage 2. The methods and the results of analysis are described in the following two sections.

Centre level correlations analysis

Analysis focussed on each OCR GCSE specification in summer 2012. For each specification, analysis first identified those centres with at least 20 candidates entering the given specification in both 2011 and 2012. For each specification, within these centres, the percentage of students achieving grade A or above and the percentage of students achieving grade C or above was calculated. The correlation between these percentages and the following two variables were then estimated:

- The percentage achieving at the relevant grade or above in the same subject[3] in summer 2011.
- The average level of key stage 2 achievement across all candidates taking the given subject in the centre in 2012.

If the correlation with the former is greater than the correlation with the latter it may indicate that historical data on achievement within centres may be a more reliable predictor of future performance than prior attainment at key stage 2.

Results for individual OCR specifications are shown in appendix B. A summary of these results across the 41 specifications included in analysis are shown in table 1. It can be seen from this table that, across specifications, the average correlations with achievement in 2011 were 0.74 and 0.66 at grades A and C respectively. In contrast, the average correlations with the mean key stage 2 achievement within centres were a little lower at 0.61 and 0.64 respectively. Close inspection of the full set of results reveals that at grade A the centre correlations with 2011 results are higher than the correlations with key stage 2 averages for all but 3 specifications. Similarly, at grade C the centre correlations with 2011 results are higher than the correlations with key stage 2 averages for all but 11 specifications. Further inspection reveals key stage 2 results are particularly likely to be the more highly correlated of the two indicators where a new specification has been introduced for a given subject. For example, both Maths and English introduced new specifications in 2012 and in both cases, there were instances where the centre

[3] But not necessarily the same specification. Analysis found that, provided a specification was available in both 2011 and 2012, the vast majority of common centres by subject were also common centres by specification. For this reason, distinguishing between centres with historical data in the same specification and centres with historical data in the same subject was not considered worthwhile.

level correlations with key stage 2 attainment were greater than the correlations with historical performance in the centre. However, for the vast majority of subjects where no new specification has been introduced, the historical performance of a centre will tend to be a better indicator of its future performance than the average prior attainment of pupils within that centre.

Table 1: Summary of correlations between centre level attainment and other centre level predictors

	Centre level correlations of…			
	% achieving grade A in 2012 with…		% achieving grade C in 2012 with…	
Summary of centre level correlations across 41 specifications	**% achieving grade A in 2011**	**Mean KS2 attainment within centre in 2012***	**% achieving grade C in 2011**	**Mean KS2 attainment within centre in 2012***
Mean	0.74	0.61	0.66	0.64
Median	0.78	0.62	0.67	0.65
Min	0.30	0.37	0.26	0.33
Max	0.91	0.84	0.85	0.89
Standard Deviation	0.13	0.12	0.12	0.12
Number of correlations below 0.3	0	0	1	0
Number of correlations 0.3-0.4	1	4	0	1
Number of correlations 0.4-0.5	2	3	1	2
Number of correlations 0.5-0.6	4	11	9	14
Number of correlations 0.6-0.7	4	12	13	9
Number of correlations 0.7-0.8	14	10	13	12
Number of correlations 0.8-0.9	15	1	4	3
Number of correlations above 0.9	1	0	0	0

* Restricted to centres with at least half of their candidates having relevant prior attainment data

Comparison of putative grade distribution with "gold standard"

At face value the results above are fairly encouraging for the use of common centres data. After all, if individual centres get a better idea of their future performance from historical results than from student prior attainment, it would appear reasonable to suggest that this is the most reliable data for awarding bodies to use to predict overall performance within a subject. However, although appealing, such logic ignores the fact that key stage 2 attainment is not applied at the level of individual centres but at the level of individual candidates. As such, although it suffers from a lower correlation with GCSE performance[4], it can benefit from being applied across a greater amount of data. That is, the common centres approach relies upon relatively high correlations but applied across perhaps a few hundred centres. In contrast, the key stage 2 prediction matrices approach relies upon lower correlations but applied across several thousand individual candidates. It is possible that the ability to apply key stage 2 prediction matrices at the candidate level may compensate for the lower correlations.

The overall relative effectiveness of key stage 2 based prediction matrices relative to the common centres approach was explored using achievement data from summer 2011 by Benton and Sutch (2012). The idea behind analysis was that an appropriate method to examine the margin of error of any method used to set grade boundaries is to compare the results from the given method to the results that would be gained if we were able to use a far more powerful variable in setting grade thresholds – namely, concurrent GCSE attainment. Analyses within the report show that, at candidate level, the correlation between concurrent GCSE attainment and the grade achieved in any individual GCSE subject is much higher (at around 0.7) than the average correlation with KS2 (at around 0.5). For this reason it is reasonable to assume that if

[4] And indeed this correlation is lower still at the candidate level. Benton and Sutch (2012) show that, at candidate level, the average correlation across subjects between KS2 attainment and GCSE grade is roughly 0.5.

this information were available at the time of standard setting (which, of course, it cannot) we would certainly prefer the use of this information to the use of KS2. Indeed predicted distributions based upon concurrent attainment are one of the key ways in which inter-board differences are ultimately evaluated post awarding. Thus we can evaluate any method used to set grade boundaries by comparing the predicted distributions produced by the given method to those predicted by candidates' concurrent GCSE attainment. The full procedure for analysis was as follows:

- Restrict data to candidates with matching concurrent GCSE attainment; that is, candidates that have taken at least 3 other GCSEs beyond the GCSE subject being studied.
- Restrict 2011 data to OCR candidates and match in data about the UMS score of candidates.
- Restrict analysis to the 52 GCSE subjects with at least 500 year 11 candidates taking the subject with OCR in 2011.
- Generate putative grade distributions for 2011 OCR candidates using historical data from 2010[5] and based on four different possible data sources:
    - Mean concurrent GCSE.
    - Key stage 2[6]. Note that since not all pupils have matched KS2 data this is a two stage procedure. First, the putative percentage is calculated for matched candidates. Next, grade boundaries on the UMS scale[7] are identified that would yield these putative grades. Finally, these grade boundaries are applied across all pupils (matched and unmatched) to yield an overall putative grade distribution.
    - Common centres. That is, results within 2011 OCR centres in the same GCSE subject in 2010. For each OCR centre, the probability of 2011 pupils achieving any grade was estimated to be equal to the percentage of pupils in the centre who achieved that grade in 2010. Since a small number of centres will not have historical information, a similar two-stage procedure was used as for key stage 2.
    - Reproducing the cumulative percentage for OCR candidates in 2010 for 2011. That is, if is 46.7 per cent of OCR year 11 Biology candidates were awarded a grade A/A* in 2010 then the putative percentage for 2011 will be exactly that (46.7).
- Compare the putative percentages from mean GCSE score to the putative percentages from the other three methods to provide an idea of the relative accuracy of each method.

By taking the predicted percentage from mean concurrent GCSE as a "gold standard", the results of this analysis allow a comparison of the relative accuracy of three potential methods that could be used to set standards; key stage 2 prediction matrices, common centres and maintaining a fixed percentage of pupils achieving each grade within an awarding body.

The results from Benton and Sutch (2012) are reproduced in table 2. The results suggest that on average the predicted grade distribution based on KS2 is closer to the predictions from concurrent attainment than either a common centres approach or simply carrying forward the percentage achieving particular grades from the previous year. However, the differences are quite slight. Looking at the difference in medians suggests that putative grade distributions based on KS2 prediction matrices will be closer to the gold standard distribution by less than 1 percentage point.

Overall therefore, analysis suggests that, despite the encouraging centre level correlations reported earlier, the common centres approach to standard setting is no improvement on the use of key stage 2 prediction matrices.

[5] Across all boards

[6] Excluding the KS2 results of pupils studying in independent schools (even if matched data was available). This restriction was imposed to ensure consistency with the approach to the use of key stage 2 prediction matrices in practice.

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

Table 2: Extent to which different methods match the predicted 2011 grade distributions generated using concurrent GCSE attainment (Reproduced from table 4 in Benton and Sutch, 2012)

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

The value of redefining "benchmark" centres

The analysis in the previous section reveals that the use of data from common centres is unlikely to provide a superior general method for setting grade boundaries than the use of key stage 2 prediction matrices. It is now of interest to explore whether the method could potentially be improved further by restricting our common centres to those that appear to have stable characteristics over time.

Benchmark centres might be defined as common centres that are evidently stable in terms of one or more of the following characteristics:

- A large cohort size entering for a given subject.
- Minimal change in the size of the cohort entering a given subject between years.
- Historically stable results in the given subject. For example, this might be evidenced by a consistent percentage of pupils gaining grade C or above or grade A or above within the centre.
- Minimal change in the prior attainment of candidates entering the given subject in successive years.

The aim of the analysis presented in this section is to explore whether there is any advantage to defining benchmark centres in any of the four ways described above. That is, if we were to define benchmark centres in any of the above four ways and then to restrict the common centres method to our newly defined benchmark centres, would the accuracy of the method improve?

Analysis was restricted to each OCR specification taken in June 2012 where there were at least 50 common centres with 2011 each of which entered at least 20 candidates. Because of the poor performance of new specifications within the correlations analysis presented earlier these were not included within analysis. This left 39 OCR specifications available for further analysis. Details of the specifications included within analysis are given in appendix B. Within each of these specifications and for various definitions of "benchmark centre" two methods were applied to evaluate the effectiveness of the definition.

Calculating the standard error of derived grade boundaries

The standard errors of derived grade boundaries quantify the precision of benchmark centres methods based upon different definitions. Lower standard errors would indicate that using benchmark centres (defined in a particular way) provided a more accurate method for setting grade boundaries than using all available common centres. Standard errors were calculated using the following method for each specification:

1. Restrict the sample to only centres that meet the given requirements for being "benchmark centres".
2. Sample centres with replacement from this data set to produce a bootstrap sample containing the same number of centres as were identified as being available for analysis in step 1.
3. Across the sampled centres calculate the overall predicted percentage to achieve grade A/C or above. That is, calculate each candidate's probability of achieving grade A/C or above as the proportion of candidates achieving this within their centre in the given subject in 2011. These estimated probabilities are then aggregated across candidates to produce the overall predicted percentage.
4. Examine the cumulative distribution of UMS scores within the sampled centres in 2012 and identify the required grade boundary on the UMS scores[8] so that the number of candidates who would be awarded an A/C or above will equal the predicted percentage in stage 3.
5. Repeat steps 2 to 4 five hundred times to estimate the standard error of the procedure.

As with any method for estimating standard errors, all else being equal, analysis based on larger sample sizes will tend to have lower standard errors than methods based on smaller sample sizes. However, the above methodology will also give lower standard errors to methods where the predicted percentage to achieve a given grade or above accurately forecasts the percentage actually achieving above a given UMS score – a possibility that is more likely if the two quantities are highly correlated. If a given methodology produces accurate forecasts of the percentage achieving above a given UMS score then, regardless of which centres are sampled within the bootstrap procedure, the same given mark will be identified as a grade boundary and the standard error of the method will be identified as being low.

Comparing derived putative grade distributions to a "gold standard" putative grade distribution

As with the analysis of 2011 achievement data described by Benton and Sutch (2012), the idea behind analysis is that an appropriate method to examine the margin of error of any method used to set grade boundaries is to compare the results from the given method to the results that would be gained if we were able to set grade thresholds using concurrent GCSE attainment. The full procedure for analysis was as follows:

- Restrict data to candidates with matching concurrent GCSE attainment; that is, candidates that have taken at least 3 other GCSEs beyond the GCSE subject being studied.
- Restrict 2012 data to OCR candidates and match in data about the UMS score of candidates.
- Generate putative grade distributions for 2012 OCR candidates using historical data from 2011[9] and based on both:
    - Mean concurrent GCSE.
    - Benchmark centres as defined by different definitions. Since not all candidates are within benchmark centres this is a two stage process. Firstly, we identify the grade boundary on the UMS score derived from the application of the benchmark centres methodology using the definition being studied. Having identified the necessary grade boundary, the putative grade distributions across the whole

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

[9] Across all boards in the subject of interest

sample is defined by the percentage of candidates achieving at or above this UMS score.

- Compare the putative percentages from mean GCSE score to the putative percentages from different benchmark centres methods to provide an idea of the relative accuracy of each method.

By taking the predicted percentage from mean concurrent GCSE as a "gold standard", the results of this analysis allow a comparison of the relative accuracy of different approaches to defining benchmark centres.

The above methods were applied to examine different ways of defining benchmarking centres in terms of absolute size, relative change in size, stability of results and stability of prior attainment. The subsequent subsections examine the relative merits of different definitions based on each of these criteria.

Definitions based on absolute entry size

The most obvious restriction to place upon those centres used within a common centres methodology is that each centre should have a reasonable number of candidates entering the subject in the baseline year; that is, the year prior to the one in which we are setting standards. Without this restriction there is the potential for predictions for the subsequent year in any centre to be based upon an insufficiently small number of cases.

Within each subject, definitions of benchmark centres as being those entering at least 1, 10, 20, 30 and 50 candidates for the given subject in summer 2011 were tried. For each definition the standard errors of grade boundaries at grade C and grade A were calculated using the methodology described earlier.

A summary of the results of this analysis across 39 specifications is shown in table 3. These results show that greater restrictions on which centres can be included as benchmark centres tend to increase the standard errors of the methodology. For example, the standard error of the C grade boundary if we use all common centres[10] is 1.8 UMS points, however, if we restrict to centres with at least 20 entrants in 2011 the standard error rises to almost 2.1 UMS points. This implies that, although the past performance of large centres would be more likely to provide meaningful indicators of future performance, any benefit of restricting analysis to larger centres is outweighed by reducing the size of the available data overall.

Table 3: Standard errors of the benchmark centres method across different definitions of such centres in terms of the minimum number of entries in summer 2011.

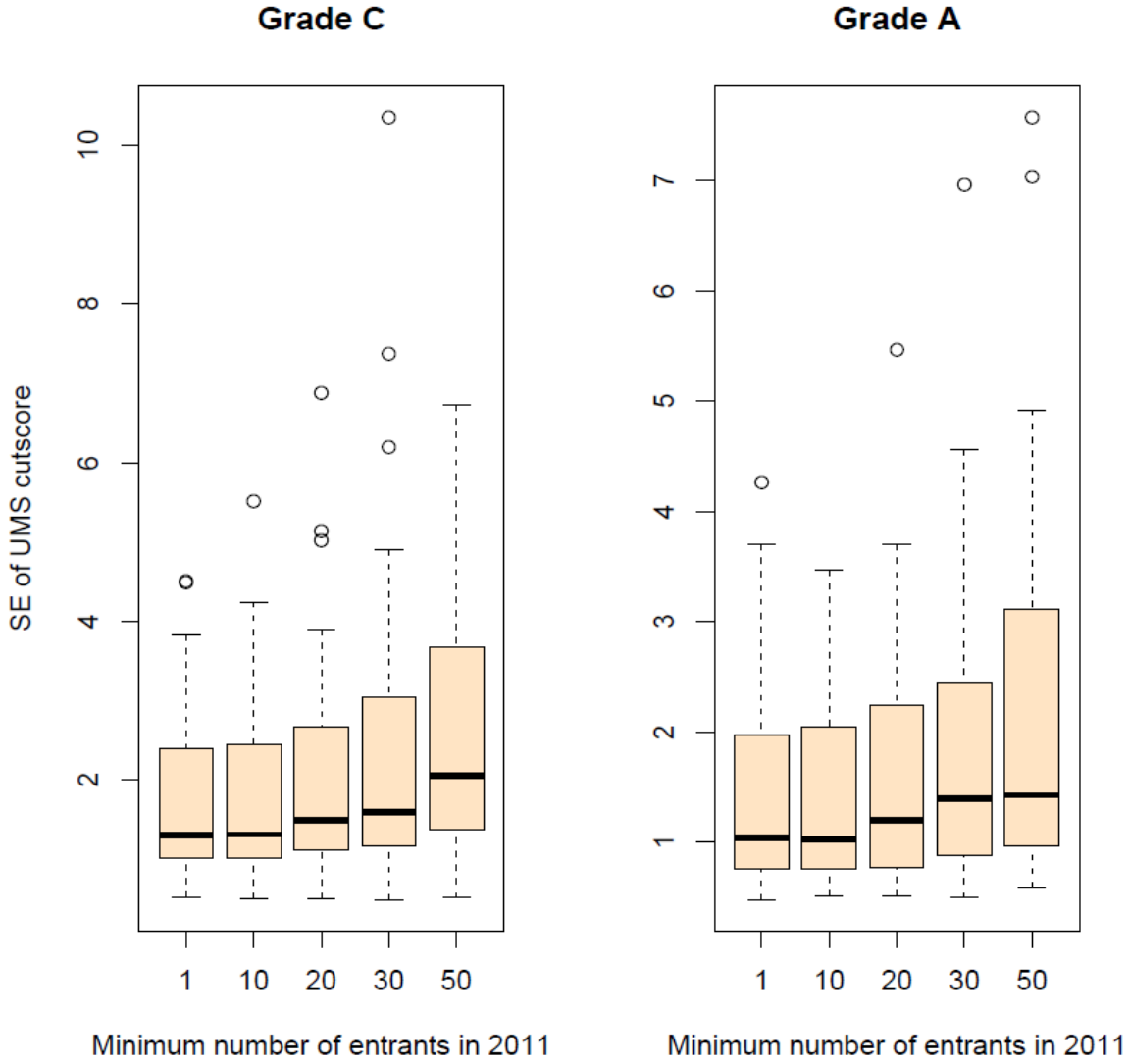
Standard errors of grade boundaries (Summary of results across 39 OCR specifications)		**Definition of benchmark centres: Minimum number of entries in summer 2011**				
		1	**10**	**20**	**30**	**50**
At grade C	**Mean**	1.80	1.81	2.07	2.46	2.65
	Median	1.31	1.32	1.50	1.60	2.06
	Standard Deviation	1.06	1.12	1.42	2.04	1.84
At grade A	**Mean**	1.43	1.41	1.63	1.86	2.30
	Median	1.03	1.02	1.20	1.39	1.42
	Standard Deviation	0.91	0.84	1.14	1.41	1.84

At both grade A and grade C it would appear that restricting benchmark centres to those with a large number of candidates has a detrimental effect on the accuracy of the method. The only exception being the possibility that there is no harm in restricting to centres with at least 10 candidates in 2011. Boxplots giving further details of these results are given in figure 1. These

[10] That is, if benchmark centres are defined as all common centres with at least 1 entrant in 2011 (that is, all of them).

confirm the conclusion that there is no benefit to restricting benchmark centres to centres of a particular size and that restricting to those of sizes greater than 20 may even be harmful.

Figure 1: Boxplots showing the relationship between minimum size of benchmark centres and the standard error of the method across different specifications.



Further analysis of the effectiveness of different definitions based on absolute entry size is given in table 4. This shows a summary of the average absolute difference between the putative grade distributions derived via a benchmark centres method and via the use of concurrent GCSE attainment. In other words this table shows how far the putative grade distribution derived via various definitions of benchmark centres differs from the “gold standard”. For example, this table shows that on average, across all 39 OCR specification used in analysis, the percentage predicted to gain grade C or above via benchmark centres using every common centre is 1.7 percentage points different from the predicted percentage from concurrent GCSE.

The results in table 4 show that restricting benchmark centres to those with a given minimum number of candidates provides little benefit in terms of how accurately the results match the gold standard putative distribution. In this way the results are consistent with the earlier findings from using bootstrapping to calculate standard errors. However, we also find that in general there is little decrease in performance associated with restricting benchmark centres to those of a particular size. This does not exactly reflect our expectations based on the analysis of standard errors. However, these differences may simply be caused by our inability to reliably capture small differences in accuracy with such a small number of observations (just 39 specifications). Furthermore, there is a little evidence of restricting benchmark centres to those with at least 50

candidates being detrimental to performance. The median difference to the gold standard distribution appears a little higher compared to having no size restrictions, and (more obviously) the mean difference is substantially higher with this last difference being driven by a large increase in the maximum.

Table 4: Summary of differences between putative cumulative grade distributions based upon concurrent GCSE attainment and benchmark centres method across different definitions of such centres in terms of the minimum number of entries in summer 2011.

Summary of absolute differences from "gold standard" putative grade distribution across 39 OCR specifications		**Definition of benchmark centres: Minimum number of entries in summer 2011**				
		1	**10**	**20**	**30**	**50**
At Grade C	**Mean**	1.7	1.6	1.7	1.6	3.9
	Median	1.3	1.2	1.3	1.1	1.5
	Min	0.2	0.2	0.0	0.2	0.1
	Max	8.2	8.2	8.2	8.2	25.2
	Standard Deviation	1.6	1.5	1.6	1.7	6.5
At Grade A	**Mean**	2.8	2.6	2.7	2.7	9.5
	Median	2.6	2.5	2.6	3.0	3.0
	Min	0.1	0.3	0.3	0.1	0.1
	Max	6.0	5.6	5.9	6.5	74.7
	Standard Deviation	1.6	1.5	1.6	1.6	20.8

Definitions based relative change in entry size

Having seen that restricting benchmark centres to those of a particular absolute size is unlikely to be of any benefit, we now turn our intention to whether benchmark centres should exclude those centres that show a large change in the number of entrants between successive years. The rationale for such a restriction might be that centres with a large change in the number of entrants might also display a large change in the standard of their entrants and so their historical performance would not provide a reliable indicator of future performance.

Within each subject, definitions of benchmark centres as being those where the number of entrants changed by no more than 30, 50, 100, and 200 per cent between summer 2011 and summer 2012[11] were tried. For each definition the standard errors of grade boundaries at grade C and grade A were calculated using the methodology described earlier.

A summary of the results of this analysis across 39 specifications is shown in table 5. These results show that reducing restrictions on which centres can be included as benchmark centres, tends to decrease the standard errors of the methodology. For example, the standard error of the C grade boundary if we restrict to common centres with a no more than 30 per cent change in size is 2.3 UMS points, however, if we allow centres with a 200 per cent change in size to be included the standard error reduces to around 1.8 UMS points. This again implies that any benefit from improved predictive validity for individual centres by restricting analysis to centres with a stable number of entrants is outweighed by reducing the overall size of the available data.

[11] The percentage change was defined as:
100*(Maximum(Entrants in 2011,Entrants in 2012)/ Minimum(Entrants in 2011,Entrants in 2012)-1).
Thus, for example, both a change from 100 candidates in 2011 to 130 in 2012 and a change from 130 candidates in 2011 to 100 candidates in 2012 would be defined as a 30 per cent change.

Table 5: Standard errors of the benchmark centres method across different definitions of such centres in terms of the change in entry size between summer 2011 and summer 2012.

Standard errors of grade boundaries (Summary of results across 39 OCR specifications)		Definition of benchmark centres: Maximum allowable percentage change in number of entrants between 2011 and 2012			
		30	50	100	200
At grade C	Mean	2.28	1.98	1.82	1.77
	Median	1.59	1.49	1.35	1.33
	Standard Deviation	1.52	1.20	1.08	0.99
At grade A	Mean	1.72	1.55	1.41	1.38
	Median	1.28	1.14	1.04	1.04
	Standard Deviation	1.11	1.00	0.87	0.83

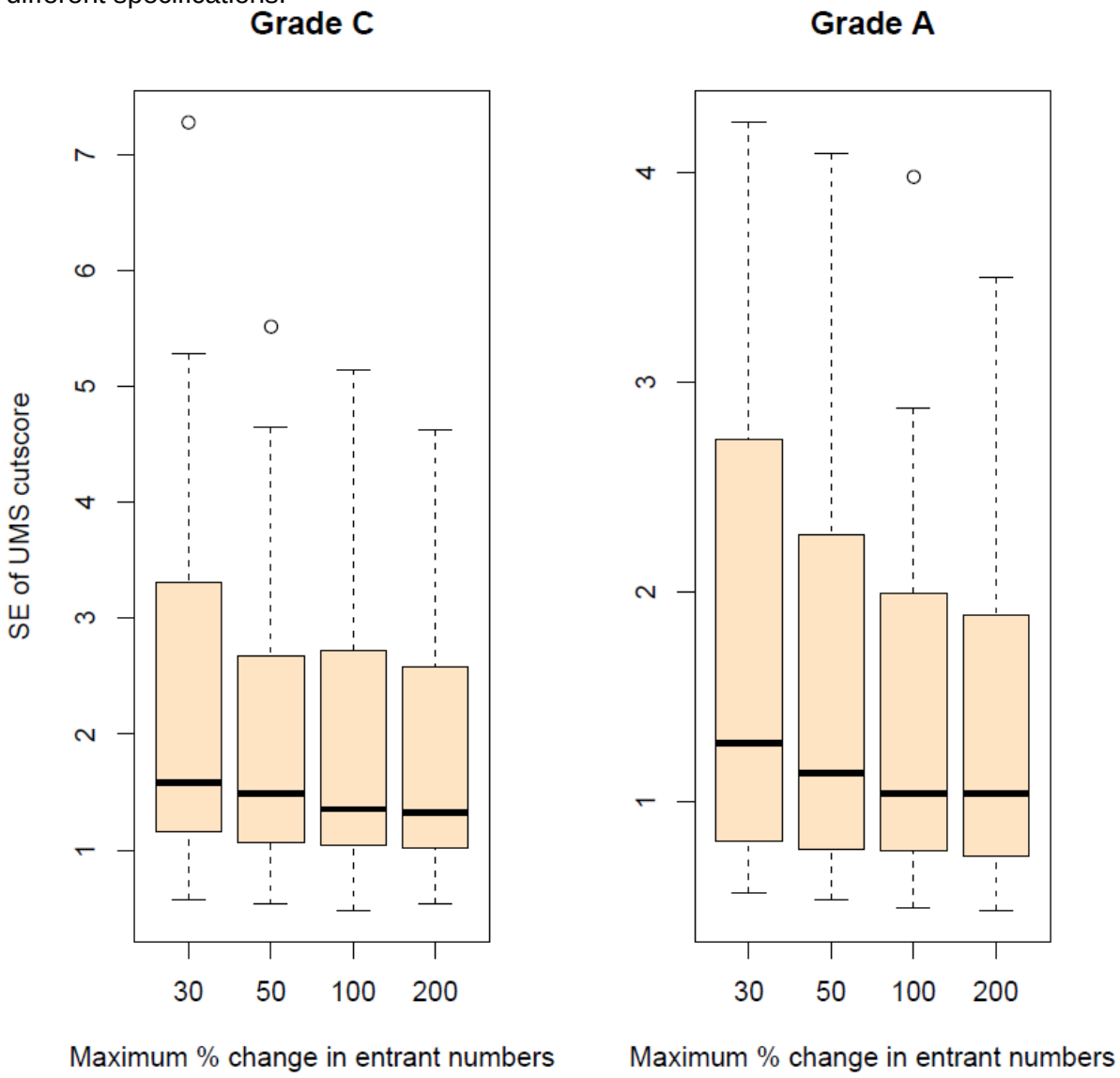
At both grade A and grade C it would appear that restricting benchmark centres to those with a stable entry has a detrimental effect on the accuracy of the method. Having said this, removing centres with a greater than 100 per cent change in their number of entrants has little impact on the accuracy of the method. Boxplots giving further details of these results are given in figure 2. These confirm the conclusion that restricting benchmark centres to those with a stable entry tends to decrease rather than increase the accuracy of the method.

Further analysis of the effectiveness of different definitions based on change in entry size, in terms of differences with gold standard putative distributions, is given in table 6. The results in table 6 show little difference in the accuracy of the method dependent upon the restrictions we place on benchmark centres. This is particularly true at grade A, although again we note this may be caused by our inability to reliably capture small differences in accuracy with just 39 observations. However, at grade C there is a little evidence that tightly restricting our definition of benchmark centres can reduce the accuracy of the method as both the mean and median differences with the gold standard increase slightly if we restrict our analysis to centres where the size of the entry has changed by less than 30 per cent.

Table 6: Summary of differences between putative cumulative grade distributions based upon concurrent GCSE attainment and benchmark centres method across different definitions of such centres in terms of the change in entry size between summer 2011 and summer 2012.

Summary of absolute differences from "gold standard" putative grade distribution across 39 OCR specifications		Definition of benchmark centres: Maximum allowable percentage change in number of entrants between 2011 and 2012			
		30	50	100	200
At Grade C	Mean	1.8	1.6	1.5	1.6
	Median	1.7	1.5	1.2	1.3
	Min	0.1	0.2	0.2	0.2
	Max	4.3	3.8	5.3	6.9
	Standard Deviation	1.1	1.1	1.2	1.3
At Grade A	Mean	2.3	2.3	2.4	2.6
	Median	2.5	2.5	2.5	2.5
	Min	0.3	0.1	0.3	0.1
	Max	6.5	7.0	5.3	5.6
	Standard Deviation	1.4	1.5	1.3	1.5

Figure 2: Boxplots showing the relationship between maximum allowable percentage change in the size of benchmark centres between years and the standard error of the method across different specifications.



Definitions based historical stability of results

Having dismissed the number of entrants within a centre as a productive means to define benchmark centres we next turn to the possibility of using historical data on individual centre performance. It would appear reasonable that if we were to restrict benchmark centres to being those common centres where a consistent percentage of pupils achieve at a given level we could be more confident about historical results being carried forward. For example, it may seem reasonable to assume that if for the past 3 years roughly 50 per cent of candidates have achieved grade C or above within a given centre, we can be confident that we would expect roughly 50 per cent to achieve grade C or above in the next year. Furthermore, we might expect to have greater confidence in this prediction than if a centre's results were unstable over the previous 3 years.

The stability of each centre's results was defined using historical data on achievement in each subject in the years 2009, 2010 and 2011. Within each centre, for each subject, at both grade A and grade C, the difference between the maximum percentage achieving the given grade or above in any of these years and the minimum percentage achieving the given grade or above in any of these years was calculated. Definitions of benchmark centres as being those where the

percentage of candidates achieving at each grade or above changed by no more than 5, 10, 15, 20 and 50 percentage points between 2009 and 2011 were tried. In each case benchmark centres were also restricted to those with at least 10 entrants in each year between 2009 and 2012[12]. For each definition the standard errors of grade boundaries at grade C and grade A were calculated using the methodology described earlier.

A summary of the results of this analysis across 39 specifications is shown in table 7. These results show that restricting benchmark centres to those with historically stable results tends to increase the standard error of the method. Indeed, it can be seen that restricting benchmark centres to those with extremely stable results can severely reduce the precision of the method. At both grade C and grade A, restricting to centres where results for the past 3 years change by no more than 5 percentage points more than doubles the standard error of the method[13]. This fact is of note because it is regularly suggested that volatility in previously stable schools' results should be used as a check on whether the correct standards have been applied within a given qualification. In particular, during the initial debate over the grades awarded for GCSE English in summer 2012 it was suggested that large changes in previously stable schools' results were clear evidence that incorrect standards had been applied. In contrast to this, our analysis suggests that restricting any such evidence to so called "stable centres" is likely to increase rather than decrease the imprecision in the standard setting process. Whilst our original analysis has shown that a common centres approach can provide a relatively accurate way of predicting future results, there is no apparent benefit from restricting this approach to stable centres. Further discussion of this issue is provided later in this report.

Finally, it is worth noting the fact that the methodology here implicitly requires restricting all analysis to centres with at least 3 years of historical data; that is, 2009, 2010 and 2011. It may be for this reason that even a very slight restriction on stability, such as requiring a change of no more than 50 percentage points in the percentage of candidates achieving a given grade or above between 2009 and 2011, leads to standard errors that are higher than those for the original common centres technique as shown in table 3.

Table 7: Standard errors of the benchmark centres method across different definitions of such centres in terms of the historical stability of results.

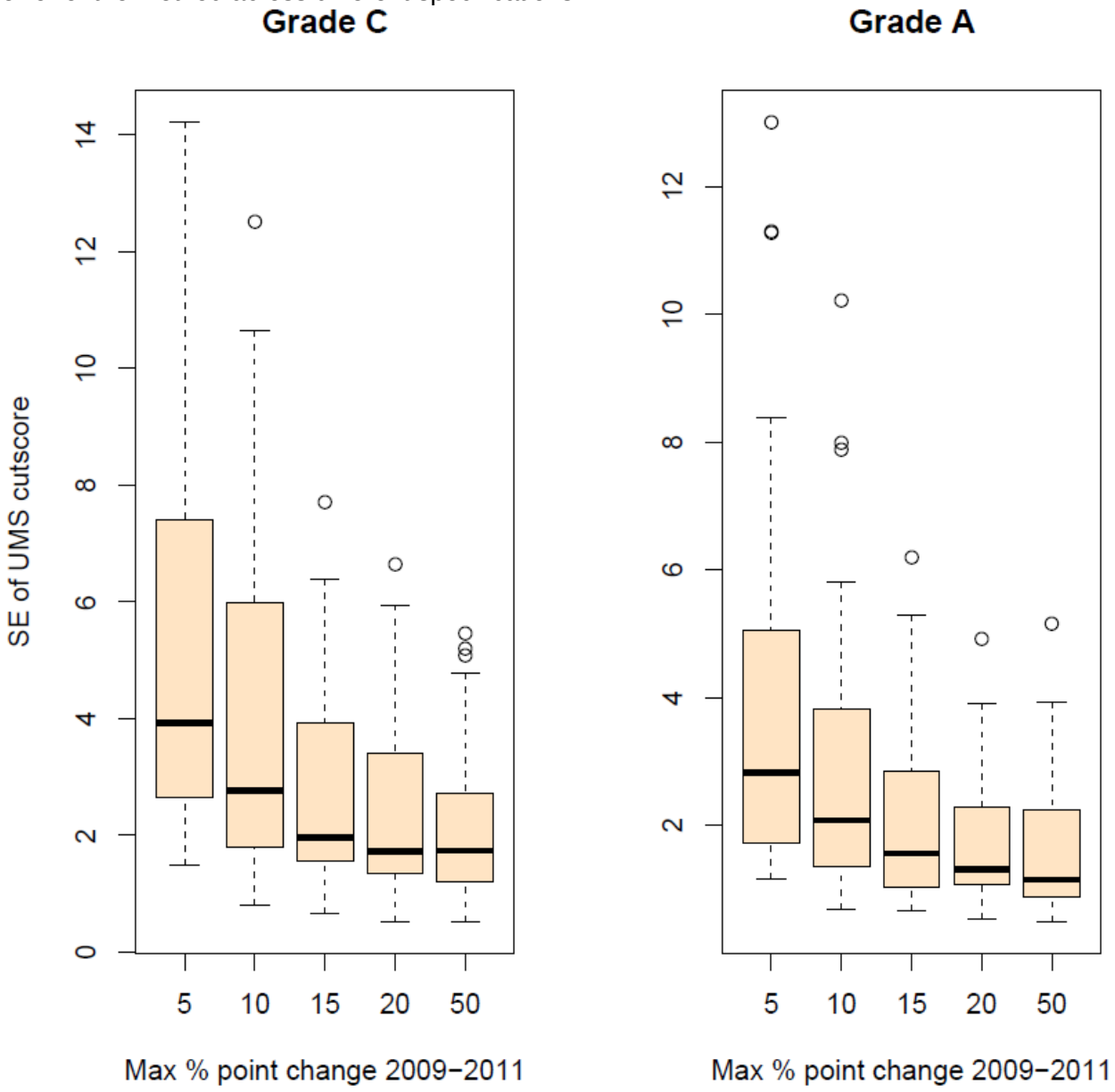
Standard errors of grade boundaries (Summary of results across 39 OCR specifications)		**Definition of benchmark centres: Maximum allowable percentage point change in percentage achieving relevant grade or above**				
		5	**10**	**15**	**20**	**50**
At grade C	**Mean**	5.15	3.73	2.81	2.52	2.13
	Median	3.93	2.77	1.96	1.72	1.74
	Standard Deviation	3.35	2.77	1.85	1.66	1.33
At grade A	**Mean**	4.11	2.90	2.09	1.80	1.64
	Median	2.83	2.08	1.56	1.31	1.15
	Standard Deviation	3.11	2.21	1.40	1.09	1.09

Boxplots giving further details of these results are given in figure 3. These confirm the conclusion that restricting benchmark centres to those with historically stable results tends to decrease rather than increase the accuracy of the method.

[12] This additional restriction was applied for two reasons. Firstly, it avoided including centres that had entered just 1 or 2 pupils each year and where all of them had achieved a given grade as benchmark centres. Secondly, earlier analysis has already shown that restricting to centres with at least 10 pupils is unlikely to harm the accuracy of the analysis and may improve the face validity of results.

[13] Compared to restricting to those with a no more than 50 percentage point range of results.

Figure 3: Boxplots showing the relationship between maximum allowable percentage point change in GCSE achievement of benchmark centres between 2009 and 2011 and the standard error of the method across different specifications.



Further analysis of the effectiveness of different definitions based on stability of historical results, in terms of differences with gold standard putative distributions, is given in table 8. At grade C the results in table 8 clearly show that placing greater restrictions on the definition of benchmark centres decreases the accuracy of the method. Both the mean and median difference from the gold standard putative grade distribution increase steadily as benchmark centres are restricted to those with ever more stable historical results. At grade A the results are less clear cut. Certainly restricting to those centres that are extremely stable (that is, exhibiting a less than 5 percentage point change over 3 years) appears to decrease the accuracy of the method. However, there is some tentative evidence to suggest that removing the most unstable centres (that is, exhibiting a greater than 50 percentage point change over 3 years) does marginally increase the accuracy of the method.

Table 8: Summary of differences between putative cumulative grade distributions based upon concurrent GCSE attainment and benchmark centres method across different definitions of such centres in terms of historical stability.

Summary of absolute differences from "gold standard" putative grade distribution across 39 OCR specifications		Definition of benchmark centres: Maximum allowable percentage point change in percentage achieving relevant grade or above				
		5	**10**	**15**	**20**	**50**
At Grade C	**Mean**	6.3	2.9	2.5	2.2	1.8
	Median	3.6	2.7	2.0	1.7	1.2
	Min	0.1	0.0	0.1	0.1	0.2
	Max	42.5	10.7	10.7	10.7	8.2
	Standard Deviation	9.7	2.6	2.0	1.9	1.7
At Grade A	**Mean**	8.8	2.7	2.1	2.4	2.8
	Median	3.2	2.3	1.8	2.1	2.7
	Min	0.1	0.0	0.1	0.1	0.3
	Max	80.7	7.1	5.6	6.6	5.8
	Standard Deviation	19.2	2.2	1.6	1.5	1.6

Definitions based stable prior attainment

Finally, we explore whether there is any benefit in restricting the common centres approach to those centres with apparently stable levels of prior attainment amongst their candidates. Such a restriction would ensure that where a centre had undergone large changes in its intake, perhaps due to converting into an academy or due to wider social changes in its local area, we would exclude them from our calculations. Such an approach would appear reasonable on the grounds that if a school has seen a large change in its intake we would not expect it to yield consistent levels of GCSE attainment between one year and the next.

Within each subject, definitions of benchmark centres as being those where the average key stage 2 attainment of candidates had changed by no more than 5, 10, 20, 50, and 100 per cent of a key stage 2 level between summer 2011 and summer 2012 were tried. For each definition the standard errors of grade boundaries at grade C and grade A were calculated using the methodology described earlier. As with the analysis looking at the value of historical stability, in each case benchmark centres were also restricted to those with at least 10 entrants in each year between 2009 and 2012[14].

A summary of the results of this analysis across 39 specifications is shown in table 9. These results again show that placing restrictions on the centres that are included within calculations tends to increase rather than decrease the imprecision of the method. Little increase in imprecision is evident from restricting to centres where the change in average key stage 2 attainment is less than 20 per cent of a level. However, it should be noted that across all specifications this restriction removes less than 30 per cent of centres as a whole[15]. Note that restricting to centres where the change in average key stage 2 attainment is less than 50 per cent of a level removes less than 5 per cent of centres as a whole which probably explains why this restriction has so little impact on results.

[14] Although for this analysis, centres that did not enter pupils every year between 2009 and 2012 were included, provided that in each year where any pupils were entered, at least 10 were entered.

[15] Compared to including all centres with key stage 2 data and at least 10 entrants each year. Calculation involves counting each centre as many times as the number of specifications for which it enters at least 10 entrants.

Table 9: Standard errors of the benchmark centres method across different definitions of such centres in terms of the change in average key stage 2 attainment of candidates between summer 2011 and summer 2012.

Standard errors of grade boundaries (Summary of results across 39 OCR specifications)		**Definition of benchmark centres: Maximum allowable change in average key stage 2 attainment within centre (percentage of a key stage 2 level)**				
		5	**10**	**20**	**50**	**100**
At grade C	**Mean**	3.29	2.41	2.09	1.88	1.91
	Median	2.54	1.77	1.38	1.38	1.39
	Standard Deviation	2.22	1.64	1.41	1.18	1.23
At grade A	**Mean**	2.41	1.86	1.62	1.52	1.49
	Median	1.85	1.40	1.14	1.03	1.04
	Standard Deviation	1.61	1.26	1.10	1.04	0.97

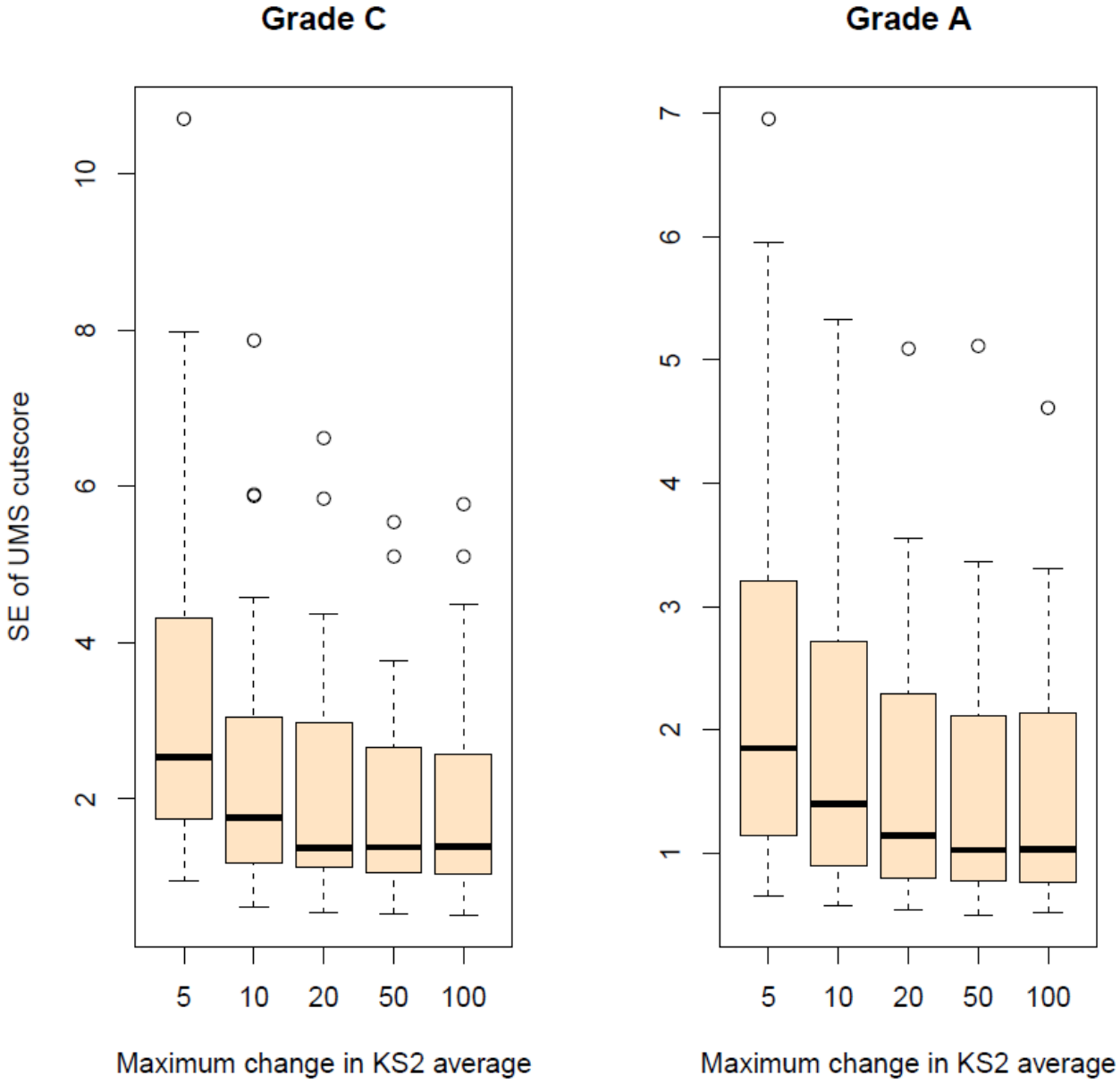
Boxplots giving further details of these results are given in figure 4. These confirm the conclusion that restricting benchmark centres to those with stable prior attainment at key stage 2 tends to decrease rather than increase the accuracy of the method.

Analysis of the effectiveness of different definitions based on stability of prior attainment, in terms of differences with gold standard putative distributions, is given in table 10. At grade C the results in table 10 match fairly closely with the results found from bootstrapping. In particular, the median difference from the gold standard putative distribution increases as we apply the tightest restrictions on the definition of benchmark centres. At grade A the results are less clear cut but it remains the case that there is no clear increase in accuracy from restricting benchmark centres to those with the least change in prior attainment between years. The very small improvements in the median for the two most restrictive categories should be treated with caution. The small number of observations available for analysis (39 specifications) means that improvements of this size cannot be seen as compelling evidence to favour either of these more restrictive definitions.

Table 10: Summary of differences between putative cumulative grade distributions based upon concurrent GCSE attainment and benchmark centres method across different definitions of such centres in terms of stability in the prior attainment of their candidates.

Summary of absolute differences from "gold standard" putative grade distribution across 39 OCR specifications		**Definition of benchmark centres: Maximum allowable change in average key stage 2 attainment within centre (percentage of a key stage 2 level)**				
		5	**10**	**20**	**50**	**100**
At Grade C	**Mean**	2.2	1.8	1.8	1.7	1.8
	Median	2.0	1.7	1.3	1.2	1.3
	Min	0.3	0.1	0.0	0.2	0.2
	Max	5.5	4.9	6.9	6.9	8.2
	Standard Deviation	1.4	1.2	1.5	1.5	1.6
At Grade A	**Mean**	2.7	2.4	2.6	2.6	2.7
	Median	2.2	2.2	2.5	2.5	2.5
	Min	0.4	0.2	0.1	0.1	0.1
	Max	6.7	4.6	5.6	5.6	5.6
	Standard Deviation	1.6	1.3	1.6	1.5	1.6

Figure 4: Boxplots showing the relationship between maximum allowable change in KS2 prior attainment within benchmark centres (quantified as a percentage of a key stage 2 level) and the standard error of the method across different specifications.

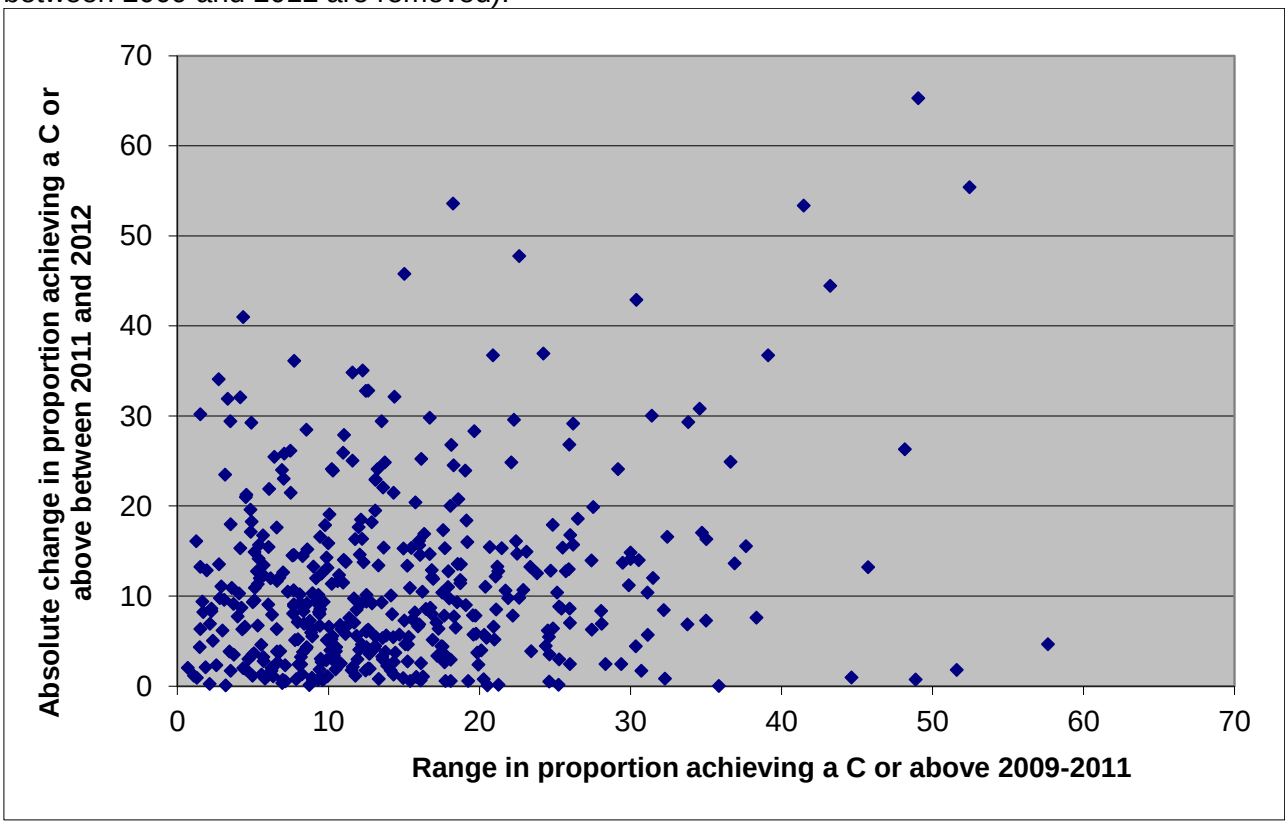


The optimal definition of a benchmark centre

Our earlier analysis has indicated that the common centres methodology can be valuable and in many cases may provide a method that is at least as accurate as the current reliance on key stage 2 results. However, our analysis indicates that there is no value in trying to restrict this method to a given set of stable "benchmark centres". Any attempt to restrict to centres that are stable in terms of their historical results, size of entry or prior attainment leads to an increase the standard errors implying that the reduced sample size is not compensated for by greater accuracy in predictions. Furthermore, there is little evidence of such restrictions leading to a closer match with the putative grade distribution based upon concurrent GCSE attainment. One reason for this is likely to be that the centres with the most stable results are often those centres that have either very high levels or very low levels of performance. As such the most stable centres may contain relatively few pupils close the grade boundary and hence provide very little useful information for the task in hand.

Another reason is that even centres that have stable results over a period of the previous 3 years do not necessarily display stable results between 2011 and 2012. An example illustrating this is shown in figure 5. The data in this figure is from OCR specification J630 (Science A); chosen because of all the specifications included within analysis it has the largest entry. Figure 5 shows the association between the range in the percentage achieving C or above between 2009 and 2011 and the change in results between 2011 and 2012. Centres where more than 95 per cent achieve a C or above in 2011 are removed from this graph as these are not very informative in the procedure for setting grade boundaries. In order to keep the graph uncluttered centres entering less than 50 candidates are also removed. Whilst there is some correlation between historical stability and stability between 2011 and 2012 it is quite small (correlation=0.21). As can be seen, even centres where the proportion achieving a C has ranged across less than 5 percentage points between 2009 and 2011 can display a sudden change in the proportion in 2012. Conversely, a number of centres where the percentage achieving a C or above has ranged by more than 20 percentage points historically can remain relatively stable between 2011 and 2012. This may explain why attempting to restrict the methodology to stable centres has no benefit in terms of its accuracy.

Figure 5: Historical stability of centres and change between 2011 and 2012 in terms of proportion of candidates achieving C or above (centres where the percentage achieving a C or above in 2011 is greater than 95% and centres with less than 50 candidates in any year between 2009 and 2012 are removed).



Using more than one year of historical data

Finally, we consider whether the accuracy of the common centre's approach would increase if, rather than using a single year's performance to predict subsequent results, predictions were made on the basis of several years' worth of historical data.

In order to explore this question the standard errors of grade boundaries were calculated using the same methodology as described in the previous section. Two very slightly different methods were explored:

- Making use of all common centres and using the percentage achieving at a given grade or above in 2011 to predict the percentage expected to achieve at a given grade or above in 2012.
- Making use of all common centres with available data for all of 2009, 2010 and 2011 and using the average percentage achieving at a given grade from 2009 to 2011 to predict the percentage expected to achieve at a given grade or above in 2012.

A summary of the results of this analysis across 39 specifications is shown in table 11. These results show little difference between using a single year's performance to predict 2012 results and using average performance across the previous three years. There are several possible reasons behind this lack of difference. One possibility is that any benefit of using several years' worth of data in terms of increasing the stability of predictions may be counteracted by the potential that slightly outdated information is being used. Another possible explanation might be that the further back in time we look at a centre's results the more likely it becomes that these results were in a different specification to the one being taken in 2012. Indeed, of the 39 OCR specifications included in analysis, none were available in summer 2009 with the exception of the science units. However, even these science units had been amended in terms of their assessment model between 2009 and 2012 and so there were no instances of complete consistency across the time period.

Table 11: Comparing the standard error of the common centres method dependent upon whether average performance across 2009-2011 or 2011 performance only is used to predict an individual centre's performance in 2012[16].

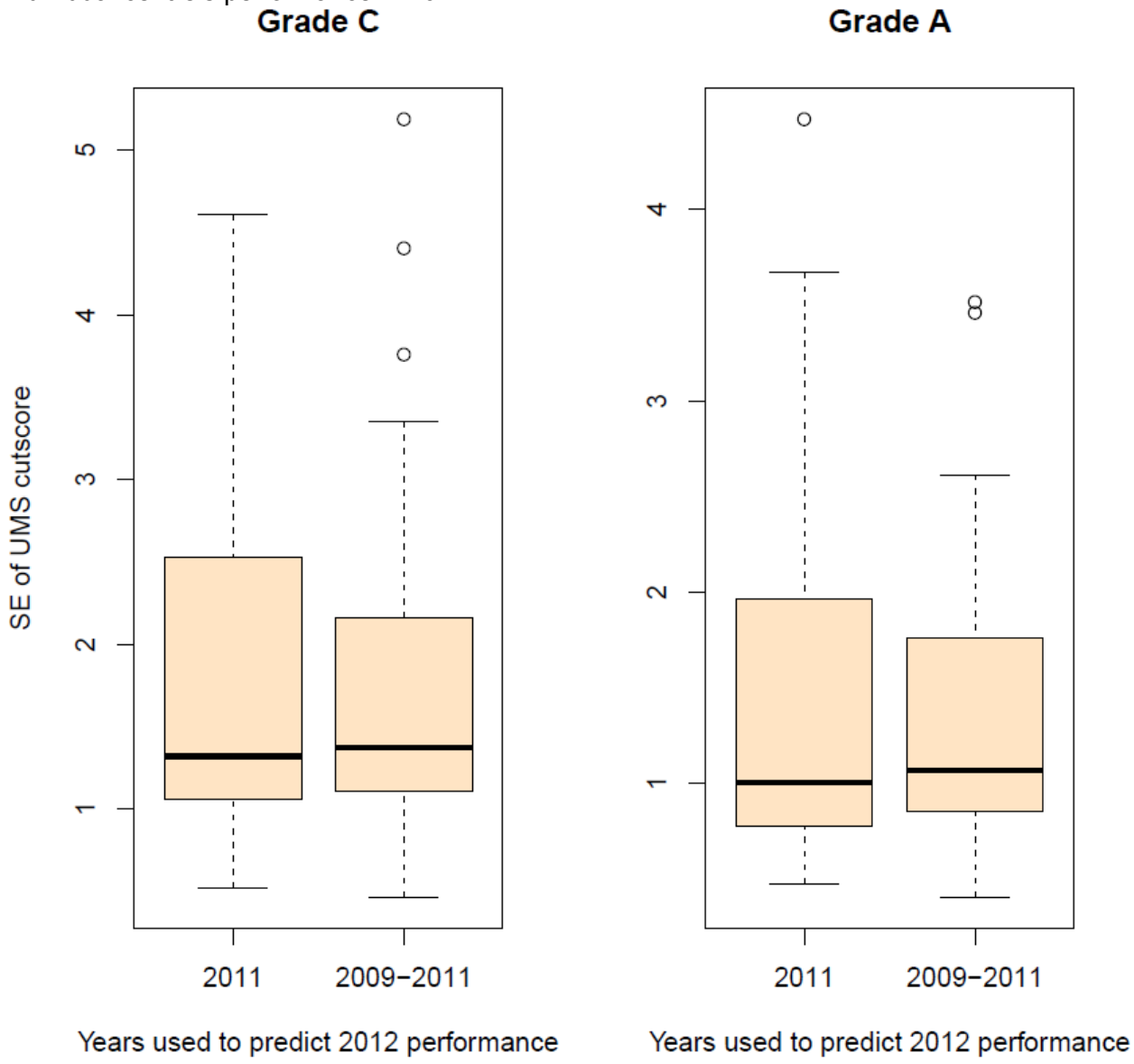
Standard errors of grade boundaries (Summary of results across 39 OCR specifications)		**Measure used to predict 2012 performance**	
		Achievement in 2011	**Average achievement 2009-2011**
At grade C	**Mean**	1.80	1.79
	Median	1.32	1.37
	Standard Deviation	1.05	1.09
At grade A	**Mean**	1.44	1.36
	Median	1.01	1.07
	Standard Deviation	0.94	0.78

Boxplots giving further details of these results are given in figure 6. These confirm the conclusion that a three year average of a centre's performance has no more predictive value than a single year's data. The results in figure 6 potentially suggest that using a three year average reduces the inter-quartile range of the inaccuracy of the common centres method[17]. However, given that this analysis explores results across just 39 specifications, it is doubtful whether this difference is statistically significant.

[16] Note that since analysis for this section was run separately from the analysis described in table 3 the results using all common centres do not match precisely.

[17] That is, it reduces the occurrence of both very high standard errors and very low standard errors.

Figure 6: Boxplots comparing the standard error of the method dependent upon whether average performance across 2009-2011 or 2011 performance only is used to predict an individual centre's performance in 2012.



Summary

- Examining the year on year correlations between centres' results show that the use of common centres to set grade boundaries has some face validity. A centre's performance in a given subject will tend to be more strongly correlated with their performance in the previous year than with the average key stage 2 attainment of the candidates entering the subject. However, the common centres approach suffers relative to the use of key stage 2 prediction matrices from the fact that predictions are usually based on results in a few hundred centres rather than being based on the achievements of several thousand individual candidates.
- More detailed analysis comparing putative grade distributions from a common centres approach to a "gold standard" based on concurrent GCSE attainment shows that the common centres approach provides a relatively accurate way of setting grade boundaries. However, analysis shows that the accuracy of this method is slightly worse than using key stage 2 prediction matrices. Having said this, the results indicate that if in particular circumstances we have a specific reason to doubt the accuracy of the key stage 2 based approach then the use of common centres may provide a reasonable alternative.
- Any attempt to refine the common centres approach by restricting to benchmark centres with particularly desirable characteristics increases the standard errors of the method. Restrictions on centres in terms of their absolute size, change in size of entry, stability of results in the given subject or stability of prior attainment are all ineffective at improving the accuracy of the method. This implies that only extreme cases should be removed from the data prior to applying the methodology. In particular it is clear that historical stability in results does not guarantee continued stability going forward and so cannot be relied upon within the process of standard maintaining.
- In addition to increasing the standard errors associated with the method, restricting benchmark centres to those with particular characteristics showed no obvious benefit in terms of producing putative grade distributions closer to those generated using concurrent GCSE attainment.
- Using centre's results in a given subject across several years historically did not prove to be more accurate than producing results based upon a single year.

Given what has been said here the straightforward conclusion would be that, if the benchmark centres methodology is to be used, the centres used within analysis should not be restricted in any way. However, it may be that from the perspective of the face validity of the method it may be desirable to remove schools where very large changes in either the size of the candidature or their prior attainment has occurred. Nonetheless, such restrictions should be limited to removing only the most extreme cases.

References

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

Appendix A: Correlations with centre level achievement for each OCR GCSE specification in summer 2012

Spec	Title	Centre level correlations of…				Number of common centres with at least 20 pupils	Number of centres with KS2 data for at least 50% of candidates
		% achieving grade A in 2012 with…		% achieving grade A in 2012 with…			
		% achieving grade A or above in 2011	Average KS2 level in 2012	% achieving grade C or above in 2011	Average KS2 level in 2012		
J160	Art and Design	0.81	0.62	0.64	0.42	137	145
J161	Art and Design: Fine Art	0.87	0.68	0.67	0.53	193	208
J253	Business Studies	0.77	0.63	0.78	0.71	171	198
J281	Latin	0.72	0.70	0.69	0.65	102	82
J302	Design and Technology: Food Technology	0.78	0.64	0.78	0.58	105	126
J303	Design and Technology: Graphics	0.85	0.68	0.76	0.67	63	101
J305	Design and Technology: Product Design	0.65	0.65	0.76	0.60	57	80
J306	Design and Technology: Resistant Materials	0.85	0.71	0.75	0.64	94	120
J315	Drama	0.60	0.57	0.55	0.51	91	113
J350*	English	0.30	0.37	0.26	0.69	117	119
J355*	English Language	0.82	0.71	0.47	0.73	262	239
J360*	English Literature	0.77	0.71	0.54	0.76	276	250
J380	Geography A	0.86	0.75	0.81	0.75	83	95
J385	Geography B	0.87	0.72	0.75	0.75	387	426
J415	History A	0.69	0.67	0.70	0.73	400	423
J417	History B	0.81	0.74	0.78	0.76	693	679
J431	Home Economics (Food and Nutrition)	0.87	0.75	0.85	0.80	74	101
J441	Home Economics (Child Development)	0.50	0.39	0.72	0.50	145	216
J526	Media Studies	0.44	0.40	0.52	0.58	108	135
J535	Music	0.82	0.74	0.82	0.71	81	133
J562*	Mathematics A	0.74	0.73	0.63	0.89	153	143
J567*	Mathematics B	0.78	0.84	0.61	0.83	189	187
J586	Physical Education	0.74	0.62	0.75	0.73	198	211
J611	Psychology	0.59	0.37	0.61	0.33	51	79
J620	Religious Studies A	0.91	0.77	0.85	0.70	51	64
J621	Religious Studies B	0.86	0.62	0.74	0.66	352	408
J630	Science A	0.70	0.52	0.61	0.72	583	608
J631	Additional Science A	0.65	0.47	0.56	0.58	577	609
J633	Biology A	0.72	0.53	0.60	0.55	474	545
J634	Chemistry A	0.70	0.53	0.59	0.54	475	541
J635	Physics A	0.72	0.53	0.56	0.53	477	542
J640	Science B	0.84	0.59	0.69	0.77	393	417
J641	Additional Science B	0.79	0.53	0.66	0.63	430	442
J643	Biology B	0.78	0.58	0.52	0.64	321	372
J644	Chemistry B	0.77	0.58	0.63	0.53	313	367
J645	Physics B	0.80	0.57	0.61	0.58	312	369
J696	Sociology	0.57	0.43	0.72	0.52	78	103
J730	French	0.81	0.66	0.71	0.66	167	149
J731	German	0.87	0.53	0.63	0.58	60	71
J732	Spanish	0.78	0.69	0.75	0.65	60	65
J938	History (Pilot)	0.46	0.50	0.58	0.45	63	68

*Specification not available in summer 2011.

Appendix B: OCR GCSE Specifications used in exploration of different definitions for benchmark centres

Specification	Title
J160	Art & Design
J161	Art & Design: Fine Art
J253	Business Studies: Single
J280	Classical Civilisation
J281	Latin
J302	Design and Technology: Food Technology
J303	Design and Technology: Graphic Products
J305	Design and Technology: Product Design
J306	Design and Technology: Resistant Materials
J307	Design and Technology: Textiles Technology
J315	Drama & Theatre Studies
J320	Economics
J380	Geography A
J385	Geography B
J415	History A (Schools' History Project)
J417	History B (Modern World)
J431	Home Economics: Food
J441	Home Economics: Child Development
J526	Media Studies
J535	Music
J586	Physical Education
J611	Psychology
J620	Religious Studies A (World Religions)
J621	Religious Studies B (Philosophy and Applied Ethics)
J630	Science A
J631	Additional Science A
J633	Biology A
J634	Chemistry A
J635	Physics A
J640	Science B
J641	Additional Science B
J643	Biology B
J644	Chemistry B
J645	Physics B
J696	Sociology
J730	French
J731	German
J732	Spanish
J938	History (Pilot)